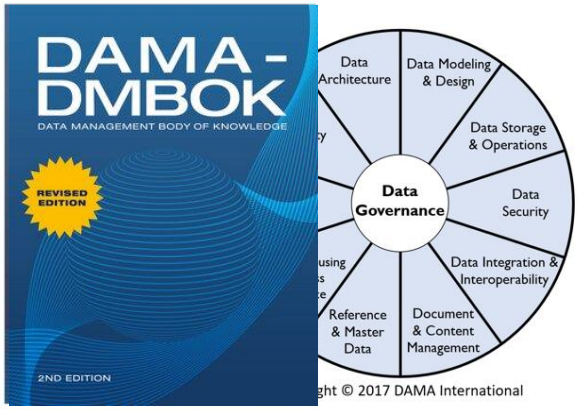


DATA MANAGEMENT for CERTIFICATION

Summary of DAMA-DMBOK2 Revised Edition® as preparation for CDMP® Exams by Veronica Diesel, CDMP Master (Data Modelling & Design, Data Governance, Data Quality), Education Director DAMA SA
(Updated September 2024)

Veronica Diesel
veronica@modelwaresystems.com

Contents

Chapter 1: Data Management

1 Introduction 19
1.1 Business Drivers 19
1.2 Goals 19
2 Essential Concepts 19
2.1 Data 19
2.2 Data and Information 19
2.3 Data as an Organisational Asset 20
2.4 Data Management Principles 20
2.5 Data Management Challenges 21
2.5.1 Data Differs from Other Assets 21
2.5.2 Data Valuation 22
2.5.3 Data Quality 22
2.5.4 Planning for Better Data 22
2.5.5 Metadata and Data Management 22
2.5.6 Data Management is Cross Functional 23
2.5.7 Establishing an Enterprise Perspective 23
2.5.8 Accounting for Other Perspectives 23
2.5.9 The Data Lifecycle 23
2.5.10 Different Types of Data 24
2.5.11 Data and Risk 24
2.5.12 Data Management and Technology 24
2.5.13 Effective Data Management Requires Leadership and Commitment 24
2.6 Data Management Strategy 25
3 Data Management Frameworks 25
3.1 Strategic Alignment Model (Henderson and Venkatraman, 1999) 25
3.2 Amsterdam Information Model (Abcouwer, Maes and Truijens, 1997) 26
3.3 The DAMA-DMBOK Framework 27
3.4 DMBOK Pyramid (Aiken) 31
3.5 DAMA Data Management Framework Evolved 31
4 DAMA and the DMBOK 34

Chapter 2: Data Handling Ethics

1 Introduction 37

2 Business Drivers 38
3 Essential Concepts 38
3.1 Ethical Principles for Data 38
3.2 Principles behind Data Privacy Law 39
3.2.1 General Data Protection Regulation of the EU 39
3.3 Online Data in an Ethical Context 40
3.4 Risks of Unethical Data Handling Practices 40
3.5 Establishing an Ethical Data Culture 40
3.6 Data Ethics and Governance 41

Chapter 3: Data Governance

1 Introduction 43
1.1 Business Drivers 43
1.2 Goals and Principles 45
1.3 Essential concepts 46
1.3.1 Data-centric Organisation 46
1.3.2 Data Governance Organisation 47
1.3.3 Data Governance Operating Model Types 48
1.3.4 Data Stewardship 49
1.3.5 Types of Data Stewards 49
1.3.6 Data Policies 50
1.3.7 Data Asset Valuation 51
2 Data Governance Activities 52
2.1 Define Data Governance for the Organisation 52
2.2 Perform Readiness Assessment 52
2.3 Perform Discovery and Business Alignment 52
2.4 Develop Organisational Touch Points 52
2.5 Develop Data Governance Strategy 53
2.6 Define the DG Operating Framework 53
2.7 Develop Goals, Principles and Policies 54
2.8 Underwrite Data Management Projects 54
2.9 Engage Change Management 55
2.10 Engage in Issue Management 55
2.11 Assess Regulatory Compliance Requirements 56
2.12 Implement Data Governance 56
2.13 Sponsor Data Standards and Procedures 57

2.14 Develop a Business Glossary 58
2.15 Coordinate with Architecture Groups 59
2.16 Define Data Asset Valuation Method 59
2.17 Embed Data Governance 59
3 Tools and Techniques 59
4 Implementation Guidelines 60
4.1 Organisation and culture 60
4.2 Adjustment and Communication 60
5 Metrics 60

Chapter 4: Data Architecture

1 Introduction 62
1.1 Business Drivers 62
1.2 Data Architecture Outcomes and Practices. 62
1.3 Essential Concepts 64
1.3.1 Enterprise Architecture Domains 64
1.3.2 Enterprise Architecture Frameworks 69
1.3.3 Enterprise Data Architecture 71
2 Activities 75
2.1 Establish Data Architecture Practice 75
2.1.1 Evaluate Existing Data Architecture Specifications 75
2.1.2 Develop a Roadmap 75
2.1.3 Manage Enterprise Requirements within Projects 76
2.2 Integrate with Enterprise Architecture 77
3 Tools 77
4 Techniques 77
4.1 Lifecycle Projections 77
4.1.1 Diagramming clarity 78
5 Implementation Guidelines 78
5.1 Readiness / Risk Assessment 78
5.2 Organisation and Cultural change 78
6 Data Architecture Governance 79
6.1 Metrics 79

Chapter 5: Data Modelling and Design

1 Introduction 81
1.1 Business Drivers 81

1.2 Goals and Principles 81
1.3 Essential Concepts 83
1.3.1 Data Modelling and Data Models 83
1.3.2 Types of Data That are Modelled 83
1.3.3 Data Model Components 84
1.3.4 Data Modelling Schemes and Notations 90
1.3.5 Data Model Levels of Detail 97
1.3.6 Normalisation 101
1.3.7 Abstraction 101
2 Activities 101
2.1 Plan for Data Modelling 101
2.2 Build the Data Model 102
2.2.1 Forward Engineering 102
2.2.2 Reverse Engineering 103
2.3 Review the Data Model 103
2.4 Maintain the Data Models 103
3 Tools 103
4 Best Practices 104
4.1 Best Practices in Naming Conventions 104
4.2 Best Practices in Database Design 104
5 Data Model Governance 104
5.1 Data Model and Design Quality Management 104
5.1.1 Develop Data Modelling and Design Standards 104
5.1.2 Review Data Model and Database Design Quality 105
5.1.3 Manage Data Model Versioning and Integration 105
5.2 Data Modelling Metrics 105
5.3 Normalisation Example (adapted from Steve Hoberman's Mastering Data Modelling Masterclass) 107

Chapter 6: Data Storage and Operations

1 Introduction 114
1.1 Business Drivers 115
1.2 Goals and principles 115
1.3 Essential Concepts 115
1.3.1 Database Terms 115
1.3.2 Data Lifecycle Management 116

1.3.3 Administrators 116
1.3.4 Database Architecture Types 117
1.3.5 Database Processing Types 120
1.3.6 Data Storage Media 121
1.3.7 Database Environments 122
1.3.8 Database Organisation 122
1.3.9 Specialised Databases 124
1.3.10 Common Database Practices 124
2 Activities 125
2.1 Manage Database Technology 125
2.1.1 Understand Database Technology 125
2.1.2 Evaluate Database Technology 125
2.1.3 Manage and Monitor Database Technology 126
2.2 Manage Database Operations 126
2.2.1 Understand Requirements 126
2.2.2 Plan for Business Continuity 126
2.2.3 Develop Database Instances 126
2.2.4 Manage Database Performance 128
2.2.5 Manage Test Data Sets 129
2.2.6 Manage Data Migration 129
3 Tools 129
4 Techniques 129
5 Implementation Guidelines 130
5.1 Readiness assessment/Risk assessment 130
5.2 Organisation and Cultural Change 130
6 Data Storage and Operations Governance 130
6.1 Metrics 130
6.2 Information asset Tracking 131
6.3 Data Audits and Data validation 131

Chapter 7 Data Security

1 Introduction 132
1.1 Business Drivers 133
1.2 Goals and Principles 134
1.3 Essential Concepts 134
1.3.1 Vulnerability 134

1.3.2 Threat 134
1.3.3 Risk 134
1.3.4 Risk Classifications: 135
1.3.5 Data Security Organisation 135
1.3.6 Security Processes 136
1.3.7 Data Integrity 136
1.3.8 Encryption 136
1.3.9 Obfuscation or Masking 136
1.3.10 Network Security Terms: 137
1.3.11 Types of Data Security 138
1.3.12 Types of Data Security Restrictions 138
1.3.13 System Security Risks 139
1.3.14 Hacking/Hacker 139
1.3.15 Social Threats to Security / Phishing 140
1.3.16 Malware 140
2 Activities 140
2.1 Identify Data Security Requirements 140
2.2 Define Data Security Policy 140
2.2.1 Security Policy contents: 140
2.3 Define Data Security Standards 141
3 Tools 142
4 Techniques 142
5 Implementation Guidelines 142
6 Data Security Governance 143
6.1 Data Security and Enterprise Architecture 143
6.2 Metrics 143

Chapter 8: Data Integration and Interoperability

1 Introduction 145
1.1 Business Drivers 146
1.2 Goals and Principles 147
1.3 Essential concepts 147
1.3.1 Extract, Transform and Load (ETL) 147
1.3.2 Latency 148
1.3.3 Replication 148
1.3.4 Archiving 149

1.3.5 Enterprise Message Format/ Canonical Model 149
1.3.6 Interaction Models 149
1.3.7 DII Architecture Concepts 149
1.3.8 Data Exchange Standards 152
2 Data Integration Activities 152
2.1 Plan and Analyse 152
2.1.1 Define data integration and lifecycle requirements: 152
2.2 Design Data Integration Solutions 153
2.2.1 Design Data Integration Architecture 153
2.2.2 Model Data Hubs, Interfaces, Messages and Data Services 153
2.2.3 Map Data Sources to Targets 153
2.2.4 Design Data Orchestration 153
2.3 Develop Data Integration Solutions 154
2.4 Implement and Monitor 154
3 Tools 154
4 Techniques 154
5 Implementation Guidelines 155
5.1 Readiness Assessment / Risk Assessment 155
5.2 Organisation and Cultural change 155
6 DII Governance 155
6.1 Data Sharing agreements 156
6.2 DII and Data Lineage 156
6.3 Data Integration metrics 156

Chapter 9: Document and Content Management

1 Introduction 157
1.1 Business Drivers 158
1.2 Goals and Principles 158
1.3 Essential Concepts 159
1.3.1 Content 159
1.3.2 Controlled Vocabularies 159
1.3.3 Documents and Records 163
1.3.4 Data Map 165
1.3.5 E-discovery 165
1.3.6 Information Architecture 166
1.3.7 Search Engine 166

1.3.8 Semantic Model 166
1.3.9 Semantic Search 166
1.3.10 Unstructured Data 166
1.3.11 Workflow 166
2 Activities 167
2.1 Plan for Lifecycle Management 167
2.1.1 Plan for Records Management 167
2.1.2 Develop a Content Strategy 167
2.1.3 Create Content Handling Policies 167
2.1.4 Define Content Information Architecture 168
2.2 Manage the Lifecycle 168
2.2.1 Capture Records and Content 168
2.2.2 Manage Versioning and Control 168
2.2.3 Backup and Recovery 169
2.2.4 Manage Retention and Disposal 169
2.2.5 Audit Documents/Records 169
2.3 Publish and Deliver Content 170
3 Tools 170
3.1 Enterprise Content Management Systems 170
3.1.1 Document Management 170
3.1.2 Content Management System 171
3.1.3 Content and Document workflow 171
3.2 Collaboration tools 171
3.3 Controlled vocabulary and Metadata tools 171
3.4 Standard Markup and exchange formats 171
3.5 E-Discovery technology 172
4 Techniques 172
5 Implementation Guidelines 172
5.1 Readiness Assessment / Risk Assessment 172
5.1.1 Records Management Maturity 172
5.1.2 E-Discovery Assessment 173
5.2 Organisation and Cultural Change 173
6 Document and Content Governance 173
6.1 Information Governance Frameworks 173
6.2 Proliferation of Information 174
6.3 Govern for Quality Content 174

6.4 Metrics 175
6.4.1 Records Management 175
6.5 E-Discovery 175
6.5.1 ECM (Electronic Content Management) 175

Chapter 10: Reference and Master Data

1 Introduction 176
1.1 Business Drivers 177
1.2 Goals and Principles 177
1.3 Essential concepts 177
1.3.1 Differences between Master and Reference Data 177
1.3.2 Reference Data 178
1.3.3 Master Data 181
1.3.4 Data Sharing Architecture 185
2 Activities 186
2.1 MDM Activities 186
2.2 Reference Data Activities 186
3 Tools and Techniques 187
4 Implementation Guidelines 187
4.1 Adhere to Master Data Architecture 187
4.2 Monitor Data Movement 187
4.3 Manage Reference Data Change 187
4.4 Data Sharing Agreements 188
5 Organisation and Cultural Change 188
6 Reference and Master Data Governance 188
6.1 Metrics 188

Chapter 11: Data Warehousing and Business Intelligence

1 Introduction 190
1.1 Business Drivers 190
1.2 Goals and Principles 191
1.3 Essential Concepts 191
1.3.1 Business Intelligence 191
1.3.2 Data Warehouse 191
1.3.3 Data Warehousing 192
1.3.4 Approaches to Data Warehousing 192
1.3.5 Corporate Information Factory (Inmon) 192

1.3.6 Dimensional DW (Kimball) 194
1.3.7 DW Architecture Components 195
1.3.8 Types of Load Processing 196
2 Activities 197
2.1 Understand Requirements 197
2.2 Define and Maintain the DW/BI Architecture 197
2.2.1 Define DW/BI Technical Architecture 197
2.2.2 Define DW/BI Management Processes 197
2.3 Develop the Data Warehouse and Data Marts 198
2.3.1 Map Sources to Targets 198
2.3.2 Remediate and Transform Data 198
2.4 Populate the Data Warehouse 198
2.5 Implement the Business Intelligence Portfolio 198
2.6 Maintain Data Products 198
3 Tools 199
3.1 Metadata Repositories 199
3.1.1 Data Dictionary / Glossary 199
3.1.2 Data and Data Model Lineage 200
3.2 Data Integration Tools 200
3.3 Business Intelligence Tools Types: 200
4 Techniques 200
5 Implementation Guidelines 201
6 DW/BI Governance 202
6.1 Enabling Business Acceptance 202
6.2 Customer / User Satisfaction 202
6.3 Service Level Agreements 202
6.4 Reporting Strategy 202
6.5 Metrics 203

Chapter 12: Metadata Management

1 Introduction 204
1.1 Business Drivers 205
1.2 Goals and Principles 205
1.3 Essential Concepts 205
1.3.1 Metadata vs. Data 205
1.3.2 Types of Metadata 206

1.3.3 ISO / IEC 11179 Metadata Registry Standard 208
1.3.4 Metadata for Unstructured data 208
1.3.5 Sources of Metadata 208
1.3.6 Types of Metadata Architecture 210
2 Activities 213
2.1 Define Metadata Strategy 213
2.2 Understand Metadata Requirements 213
2.3 Define Metadata Architecture 213
2.3.1 Create MetaModel 213
2.3.2 Apply Metadata Standards 214
2.3.3 Manage Metadata Stores 214
2.4 Create and Maintain Metadata 215
2.4.1 Integrate Metadata 215
2.4.2 Distribute and Deliver Metadata 215
2.5 Query, Report and Analyse Metadata 216
3 Tools 216
4 Techniques 216
4.1 Data Lineage and Impact Analysis 216
4.2 Metadata for Big Data Ingest 217
5 Implementation Guidelines 217
5.1 Readiness assessment / Risk Assessment 217
5.2 Organisational and Cultural Change 218
6 Metadata Governance 218
6.1 Process Controls 218
6.2 Documentation of Metadata Solutions 218
6.3 Metadata Standards and Guidelines 218
6.4 Metrics 219

Chapter 13: Data Quality Management

1 Introduction 220
1.1 Business Drivers 221
1.2 Goals and Principles 222
1.3 Essential Concepts 222
1.3.1 Critical Data 222
1.3.2 Data Quality Dimensions 223
1.3.3 Data Quality Business Rules 229

1.3.4 Data Quality Improvement Lifecycle 230
1.3.5 Common Causes of Data Quality Issues 231
2 Activities 233
2.1 Define a Data Quality Framework 233
2.2 Define High Quality Data 233
2.3 Identify Dimensions and Supporting Business rules 234
2.4 Perform an Initial Data Quality Assessment 234
2.5 Identify and Prioritise Potential Improvements 235
2.6 Define Goals for Data Quality Improvement 235
2.7 Develop and Deploy Data Quality Operations 235
2.7.1 Manage Data Quality Rules 235
2.7.2 Measure and Monitor Data Quality 236
2.7.3 Develop Procedures for Managing Data Issues 237
2.7.4 Establish Data Quality Service Level Agreements 238
2.7.5 Data Quality Response 238
3 Tools 238
3.1 Data Profiling Tools: 238
3.2 Business Rule Templates and Engines 238
3.3 Data Parsing and Formatting 238
3.4 Data Transformation and Standardisation 238
3.5 Data Enrichment 238
3.6 Incident Management System 239
4 Techniques 239
4.1 Effective Metrics 239
4.2 Data Profiling 240
4.3 Preventative Actions 241
4.4 Effective Data Quality Metrics 241
4.5 Root cause Analysis 241
4.6 Corrective Actions 242
5 Implementation Guidelines 242
5.1 Readiness Assessment / Risk Assessment 242
5.2 Organisation and Cultural Change 242
6 Data Quality and the other Knowledge Areas 243
6.1 Data Quality and Data Modelling and Design 243
6.2 Data Quality and Metadata Management 243
6.3 Data Quality and Master and Reference Data Management 243

6.4 Data Quality and Data Integration and Interoperability 243
6.5 Data Quality and Data Governance 243
8 Data Quality Appendix 244
8.1 Data Quality ISO Standard 244
8.2 Further Reading on Data Quality Dimensions 244
8.3 Statistical Process Control 246

Chapter 14: Big Data and Data Science

1 Introduction 248
1.1 Business Drivers 248
1.2 Principles 249
1.3 Essential Concepts 249
1.3.1 Data Science 249
1.3.2 The Data Science Process 250
1.3.3 Big Data 251
1.3.4 Big Data Architecture Components 252
1.3.5 Sources of Big Data 252
1.3.6 Data Lake 252
1.3.7 Services-Based Architecture 253
1.3.8 Machine Learning 253
1.3.9 Sentiment Analysis 253
1.3.10 Data and Text Mining 254
1.3.11 Predictive Analytics 254
1.3.12 Prescriptive Analytics 254
1.3.13 Unstructured Data Analytics 254
1.3.14 Operational Analytics 254
1.3.15 Data Visualisation 254
1.3.16 Data Mashups 254
2 Activities 254
2.1 Define Big Data Strategy and Business Needs 254
2.2 Choose Data Sources 255
2.3 Acquire and Ingest Data Sources 255
2.4 Develop Data Hypotheses and Methods 255
2.5 Integrate / Align Data for Analysis 255
2.6 Explore Data Using Models 255
2.7 Deploy and Monitor 255

3 Tools 256
3.1 MPP (Massively Parallel Processing) Shared-nothing Technologies and Architecture 256
3.2 Distributed File-based Databases: 256
3.3 In-database Algorithms 256
3.4 Big Data Cloud Solutions 256
3.5 Statistical computing and graphical languages 256
3.6 Data Visualisation Tools 257
4 Techniques 257
5 Implementation Guidelines 257
5.1 Strategy Alignment 257
5.2 Readiness Assessment / Risk Assessment 257
5.3 Organisation and Cultural Change 257
6 Big Data and Data Science Governance 258
6.1 Visualisation Channels Management 258
6.2 Data Science and Visualisation Standards 258
6.3 Data Security 258
6.4 Metadata 258
6.5 Data Quality 258
6.6 Metrics 258

Chapter 15: Data Management Maturity Assessment

1 Introduction 260
1.1 Business Drivers 260
1.2 Goals and Principles 262
1.3 Essential Concepts 262
1.3.1 Assessment Levels and Characteristics 262
1.3.2 Assessment Criteria 263
1.3.3 Existing DMMA Frameworks 264
2 Activities 265
2.1 Plan Assessment Activities 265
2.1.1 Define Objectives 265
2.1.2 Choose a Framework 265
2.1.3 Define Organisational Scope 265
2.1.4 Define Interaction Approach 265
2.1.5 Plan Communications 265
2.2 Perform Maturity Assessment 265

2.3 Interpret Results 266
2.4 Create a Targeted Program for Improvements 266
2.5 Re-assess Maturity 266
3 Tools 266
4 Techniques 267
4.1 Selecting a DMM Framework 267
4.2 DAMA-DMBOK Framework Use 267
5 Guidelines for a DMMA 268
5.1 Readiness Assessment / Risk Assessment 268
5.2 Organisational and Cultural change 268
6 Maturity Management Governance 268
6.1 DMMA Process Oversight 268
6.2 Metrics 268

Chapter 16: Data Management Organisation and Role Expectations

1 Introduction 271
2 Understanding Existing Organisation and Cultural Norms 271
3 Data Management Organisational Constructs 272
3.1 Decentralised Operating Model 272
3.2 Network Operating Model 273
3.3 Centralised Operating Model 273
3.4 Hybrid Operating Model 274
3.5 Federated Operating Model 275
3.6 Identifying the Best Model for an Organisation 275
3.7 DMO Alternatives and Design Considerations 276
4 Critical Success Factors 276
5 Build the Data Management Organisation 277
5.1 Identify Current Data Management Participants 277
5.2 Identify Committee Participants 277
5.3 Identify and Analyse Stakeholders 277
5.4 Involve the Stakeholders 278
6 Interactions Between the DMO and other Data-oriented Bodies 278
7 Data management Roles 279
7.1 Organisational roles 279
7.2 Individual Roles 280

Chapter 17: Data Management & Organisational Change Management

1 Introduction 283
2 Laws of Change 284
3 Not Managing a Change: Managing a Transition 285
4 Kotter's Eight Errors of Change Management 286
4.1 Error #1: Allowing Too Much Complacency 287
4.2 Error #2: Failing to Create a Sufficiently Powerful Guiding Coalition 287
4.3 Error #3: Underestimating the Power of Vision 287
4.4 Error #4: Under Communicating the Vision by a Factor of 10, 100 or 1000 287
4.5 Error #5: Permitting Obstacles to Block the Vision 287
4.6 Error #6: Failing to Create Short-Term Wins 287
4.7 Error #7: Declaring Victory Too Soon 287
4.8 Error #8: Neglecting to Anchor Changes Firmly in the Corporate Culture 288
5 Kotter's Eight Stage Process for Major Change 288
5.1 Establishing a Sense of Urgency 289
5.1.1 Kotter Identifies 9 Sources of Complacency: 289
5.2 The Guiding Coalition 290
5.3 Developing a Vision and Strategy 291
5.3.1 Why Vision is Essential 291
5.3.2 The Nature of an Effective Vision 292
5.3.3 Creating an Effective Vision (Kotter): 292
5.4 Communicating the Change Vision 293
6 The Formula for Change 293
7 Diffusion of Innovations and Sustaining Change 293
7.1 Challenges to Overcome as Innovations Spread 294
7.2 Key Elements in the Diffusion of Innovation 295
7.3 The Five Stages of Adoption 295
7.4 Factors Affecting Acceptance or Rejection of an Innovation or Change 295
8 Sustaining Change 296
8.1 Sense of Urgency/Dissatisfaction 296
8.2 Framing the Vision 296
8.3 The Guiding Coalition 296
8.3.1 Relative Advantage and Observability 296
9 Communicating the Data Management Value 296

9.1 Communication Principles 296
9.2 Audience Evaluation and Preparation 297
9.3 The Human Element 297
9.4 Communication plan 298
9.5 Keep Communicating 298
10 Metrics (Not in Revised Version) 298
11 DM Org and Change Management and Governance 299

Appendix: Possible DMBOK Version 1 Questions and Notes

1 Data Management 300
1.1 According to the DAMA DMBoK, what parts of the Data Lifecycle are integral parts of the SDLC? 300
1.2 The DMBoK identifies which of the following as common stages in the lifecycle of the information asset? 300
1.3 According to the DMBOK, which of the following are included as Data Management goals? 300
1.4 In the data lifecycle, the DMBoK recommends minimizing ROT data. ROT stands for: 301
2 Data Governance 301
2.1 According to the DAMA DMBoK, the Data Governance Council (DGC) is the highest-authority organization for Data Governance in an organization. Who should typically chair this Council? 301
3 Data Architecture 302
3.1 DMBOK Version 1 identifies the SAM and the highest level model. The second layers are conceptual views of each subject area. 302
4 Data Modelling and Design 302
4.1 What is the difference between cardinality rules and data integrity rules? 302
4.2 What is the definition of cardinality? 302
5 Document and Content Management 304
5.1 Non value-added information is often not removed because: 304
5.2 Which of the following is a reason why organizations do not dispose of non-value-adding information? 304
6 Data Warehousing and BI 305
6.1 A comparatively new architectural approach is where volatile data is provisioned in a data warehouse structure to provide transactional systems with a combination of historical and near real time data to meet customer needs. This is a definition of: 305
7 Metadata Management 306
Industry / Consensus and International Meta-data Standards: 306
7.1 What is the difference between an industry and a consensus meta-data standard? 306

Metadata Type Questions:.....306

7.2 What type of Meta-Data provides developers and administrators with knowledge and information about systems?.....306

7.3 These are examples of which type of Meta-Data: Data Stores & Data Involved, Government/ Regulatory Bodies; Roles & Responsibilities; Process Dependencies and Decomposition:.....306

8 Data Quality Management.....309

8.1 Top-down and Bottom-up data analysis and profiling are best done in concert because:.....309

8.2 The Data Quality Management cycle has four stages. Three are Plan, Monitor, and Act. What is the fourth stage?.....310

Chapter 1
Data Management

1 Introduction

Data Management is the development, execution and supervision of plans, policies, programs and practices that deliver, control, protect and enhance the value of data and information assets throughout their lifecycles.

A **Data Management Professional** is any person who works in any facet of data management, from highly technical to strategic business, to meet strategic organisational goals.

Data Management requires technical and non-technical skills, and business and IT people must collaborate to share responsibility for managing data. Data and information are vital to the operations of most organisations.

1.1 Business Drivers

- Competitive advantage of better decisions based on reliable, high quality data
- Failure to manage data is the same as the failure to manage capital
- The primary driver is to enable organisations to get value from their data assets

1.2 Goals

Data management goals within an organisation:

- Understanding and supporting the information needs of the enterprise including all stakeholders
- Capturing, storing, protecting and ensuring the integrity of data assets
- Ensuring quality of data and information
- Ensuring privacy and confidentiality of stakeholder data
- Preventing unauthorised or inappropriate access, manipulation or use of data
- Ensuring data can be used effectively to add value to the organisation

2 Essential Concepts

2.1 Data

Definitions of data:

- IT: Information that has been stored in digital form
- Facts which can be aggregated, analysed and used to make a profit, improve health or influence public policy
- Data is a means of representation which stands for things other than itself (Chisholm, 2010)
- Data is both an interpretation of the objects it represents and an object that must be interpreted (Sebastian-Coleman, 2013)
- To interpret data, context or Metadata is needed

2.2 Data and Information

- Data does not simply exist; it has to be created.
- Knowledge is required to create data in the first place

- Data is a form of information and information is a form of data
- Organisations may differentiate between information and data for communication between different stakeholders
- The terms “data” and “information” are used interchangeably in the DMBOK

2.3 Data as an Organisational Asset

An asset is an economic resource, that can be owned or controlled, and that holds or produces value. As organisations become more reliant on data to make decisions the value of data assets is more clearly established.

2.4 Data Management Principles

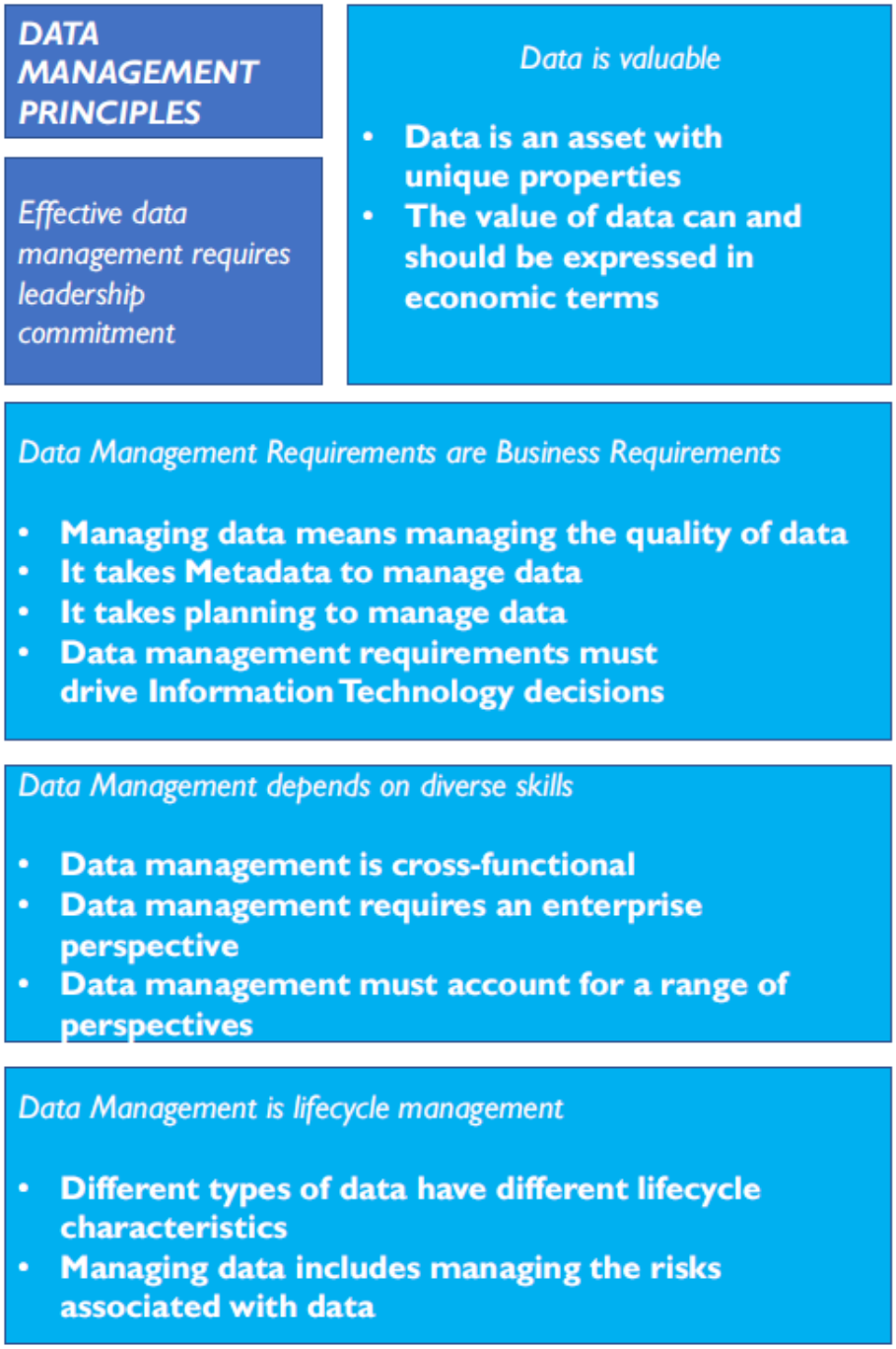

DATA MANAGEMENT PRINCIPLES
Effective data management requires leadership commitment
Data is valuable
Data is an asset with unique properties
The value of data can and should be expressed in economic terms
Data Management Requirements are Business Requirements
Managing data means managing the quality of data
It takes Metadata to manage data
It takes planning to manage data
Data management requirements must drive Information Technology decisions
Data Management depends on diverse skills
Data management is cross-functional
Data management requires an enterprise perspective
Data management must account for a range of perspectives
Data Management is lifecycle management
Different types of data have different lifecycle characteristics
Managing data includes managing the risks associated with data

Figure 1 Data Management Principles

- **Data is an asset with unique properties**: Not consumed as it is used unlike financial or physical assets.
- **The value of data can and should be expressed in economic terms**: No standards yet. Organisations should develop consistent ways to quantify value. Measure costs of low quality as well as benefits of high quality data.
- **Managing data means managing the quality of data**: Understand stakeholders' requirements for quality and ensure data is fit for their purpose.
- **It takes Metadata to manage data**: Metadata enables us to understand what intangible data is, and how to use it. Metadata originates from processes related to data creation, processing and use such as architecture, modelling, stewardship, governance, DQ management, SDLC, IT, business operations and analytics.
- **It takes planning to manage data**: Complex tech and business process landscapes, with moving data, coordinating work and keeping results aligned requires planning (architectural and business process)
- **Data management is cross functional; it requires a range of skills and expertise**: Collaboration between technical and business skills.
- **Data management requires an enterprise perspective**: To be effective. That is why DM and DG are intertwined.
- **Data management must account for a range of perspectives**: DM must constantly evolve to keep up with the ways data is created and used.
- **Data management is lifecycle management**: DM practices need to account for the data lifecycle.
- **Different types of data have different lifecycle characteristics**: DM practices need to recognise this and be flexible enough to meet different kinds of lifecycle requirements.
- **Managing data includes managing the risks associated with data**: Lost, stolen, misused and ethical implications. Managed as part of the lifecycle.
- **Data management requirements must drive Information Technology decisions**: Technology serves, rather than drives an organisation's strategic data needs.
- **Effective data management requires leadership commitment**: Complex processes of DM require the vision and purpose that comes from committed leadership.

2.5 Data Management Challenges

2.5.1 Data Differs from Other Assets

Other assets exist in one place at one time. Data is different:

- Data is not tangible.
- It is durable and does not wear out.
- The value changes as it ages.
- Easy to copy and transport.
- Not easy to reproduce if lost or destroyed.
- can be stolen but not gone.
- Not consumed when used.
- The same data can be used by multiple people at the same time.

Data has a value but it is challenging to measure it. It must be managed with care.

2.5.2 Data Valuation

Value is the difference between the cost of a thing and the benefit derived from that thing. The value of data is contextual and often temporal.

Sample cost categories:

- Obtaining and storing data
- Replacing lost data
- Impact of missing data to the organisation.
- Risk mitigation and potential costs of risks associated with data.
- cost of improving data.
- Benefits of higher quality data.
- What competitors would pay for data / what data could be sold for.
- expected revenue from innovative uses of data.

2.5.3 Data Quality

Define business needs and the characteristics that make data high quality with business data consumers. People who use data need to assume data is reliable and trustworthy as they need to use the data to learn and create value, and make business decisions. Poor quality data is costly to the organisation.

Costs of poor quality data:

- Scrap and rework.
- Work-arounds and hidden correction processes.
- Organisational inefficiencies or low productivity.
- Organisational conflict.
- Low job satisfaction.
- Opportunity costs, including inability to innovate.
- Compliance costs or fines.
- Reputational costs.

Benefits of high quality data:

- Improved customer experience.
- Higher productivity.
- Reduced risk.
- Ability to act on opportunities.
- Increased revenue.
- Competitive advantage gained from insights on customers, products, processes and opportunities.

2.5.4 Planning for Better Data

View data as a product and plan for quality throughout its lifecycle. Planning is a collaboration between business and IT as it involves architecture, modelling and other design functions.

2.5.5 Metadata and Data Management

Includes:

- Business Metadata (Ch 12)
- Technical Metadata (Ch 12)

- Operational Metadata (Ch 12)
- Metadata embedded in: (Ch 4-11)
  - data architecture
  - data models
  - Data Security requirements
  - Data integration standards
  - Data operational processes

2.5.6 Data Management is Cross Functional

Data is managed in different places within the organisation by teams that have responsibility for different phases of the lifecycle.

2.5.7 Establishing an Enterprise Perspective

Data is a 'horizontal' as it moves across all the 'verticals' (sales, marketing, operations) of the organisation. Stakeholders assume the organisation's data should be coherent, and the goal of managing data is the make the disparate data originating from different places fit together in common sense ways so that it is usable by a wide range of consumers.

2.5.8 Accounting for Other Perspectives

People who create data must be aware others will use it later. Plan around other potential uses of the data to account for legal and compliance regulations and prevent future misuse.

2.5.9 The Data Lifecycle

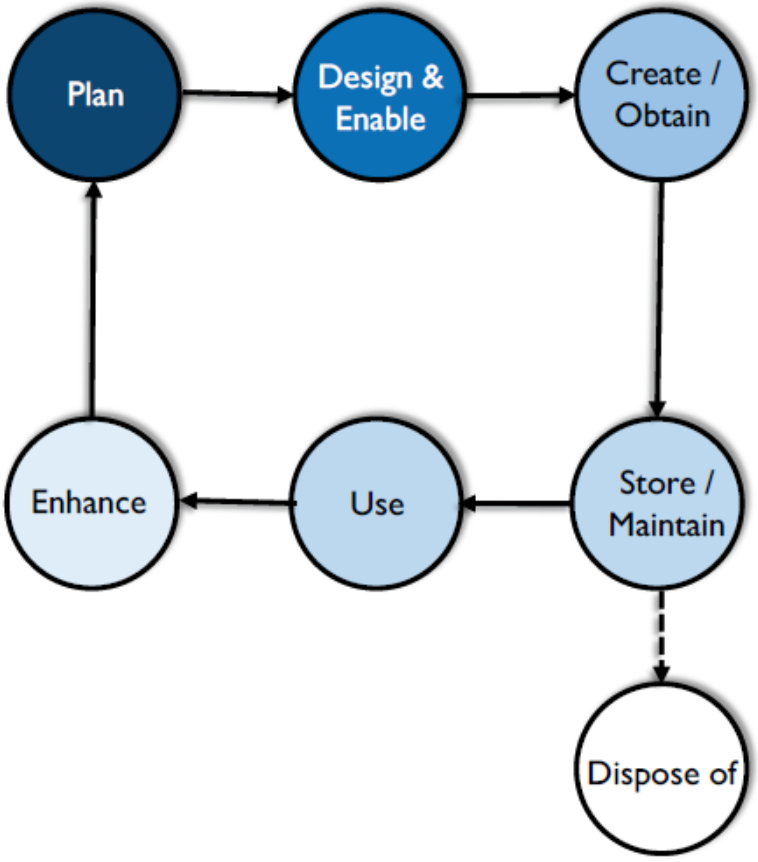


Figure 2 Data Lifecycle Key Activities

The data lifecycle is based on the product lifecycle. Throughout its lifecycle data may be cleansed, transformed, merged, enhanced or aggregate, and new data may be created. Managing data involves interconnected processes aligned with the lifecycle.

Data also has lineage, from point of origin to point of usage, which must be documented.

Implications of the focus of data management on the data lifecycle:

- **Creation and usage are the most critical points in the data** lifecycle: it costs money to produce data and it only has value when it is used.
- **Data Quality must be managed throughout the** lifecycle (see Chapter 13).
- **Metadata Quality must be managed through the data lifecycle** in the same way as the quality of other data (See Chapter 12).
- **Data Security must be managed throughout the data lifecycle:** Protection from creation to disposal (see Chapter 7).
- **Data management efforts should focus on the most critical data.**

2.5.10 Different Types of Data

Classify the data to be managed as different types require different processes. Tools of data management focus on classification and control.

- By type:
  - Transactional data
  - Refence Data
  - Master Data
  - Metadata
  - Category data
  - Resource data
  - Detailed transaction data
- By content:
  - Data domains
  - Subject areas
- By format
- By the level of protection which the data requires

2.5.11 Data and Risk

- Low quality data represents risk as it is not right.
- Data may be misunderstood and misused.
- Information gaps
- Data privacy regulations
- Stakeholders requiring privacy may be broader than previously thought.

2.5.12 Data Management and Technology

Data requirements aligned with business should drive decisions about technology.

2.5.13 Effective Data Management Requires Leadership and Commitment

A Chief Data Officer (CDO) leads data management initiatives and ensures data management is business driven instead of IT driven. The CDO leads cultural change required for the organisation to have a more strategic approach to its data.

2.6 Data Management Strategy

A strategic plan is a high-level course of action to achieve high level goals. Should address all knowledge areas of the DAMA-DMBOK.

Components of a data management strategy:

- A compelling vision for data management
- A summary business case for data management with examples
- Guiding principles, values and management perspectives
- The mission and long term directional goals of data management
- Proposed measures of data management success
- Short term (12-24 months) Data Management program objectives that are SMART (Specific, Measurable, Actionable, Realistic, Time-bound)
- Description of Data Management roles, organisations and their responsibilities
- Descriptions of the Data Management program components and initiatives
- Prioritised program of work with scope boundaries
- Draft implementation roadmap with projects and action items

Deliverables:

- **Data Management Charter:** Overall vision, business case, goals, principles, measures of success, critical success factors, risks, operating model
- **Data Management Scope Statement:** Goals and objectives for usually 3 years, roles, organisations and individual leaders accountable
- **Data Management Implementation Roadmap:** Specific programs, projects, tasks and milestones

3 Data Management Frameworks

A framework helps to understand data management comprehensively and see the relationships between component pieces. Frameworks are developed at different levels of abstraction and provide a range of perspectives.

Five models are presented in the DMBOK:

- The **Strategic Alignment Model** and the **Amsterdam Information Model** show high level relationships that influence how the organisation manages data.
- The **DAMA DMBOK Framework** (DAMA Wheel, Hexagon and Context Diagram) describes Data Management Knowledge Areas and explains their visual representation within the DMBOK.
- The final two are rearrangements of the DAMA Wheel.

3.1 Strategic Alignment Model (Henderson and Venkatraman, 1999)

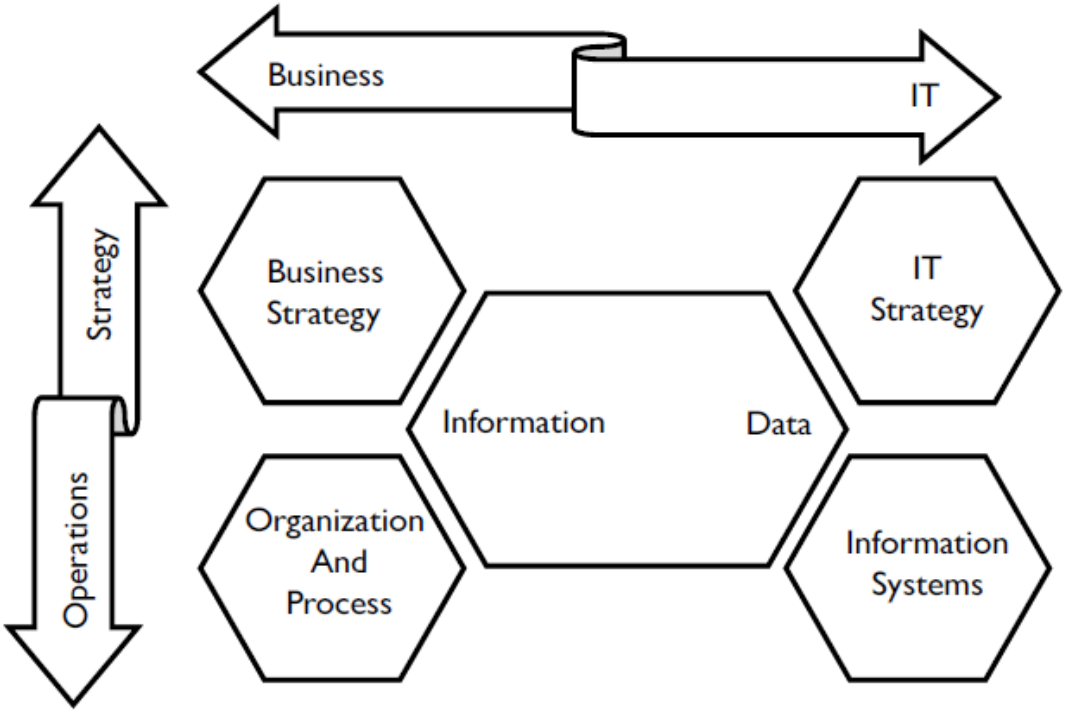


Figure 3 Strategic Alignment Model

3.2 Amsterdam Information Model (Abcouwer, Maes and Truijens, 1997)

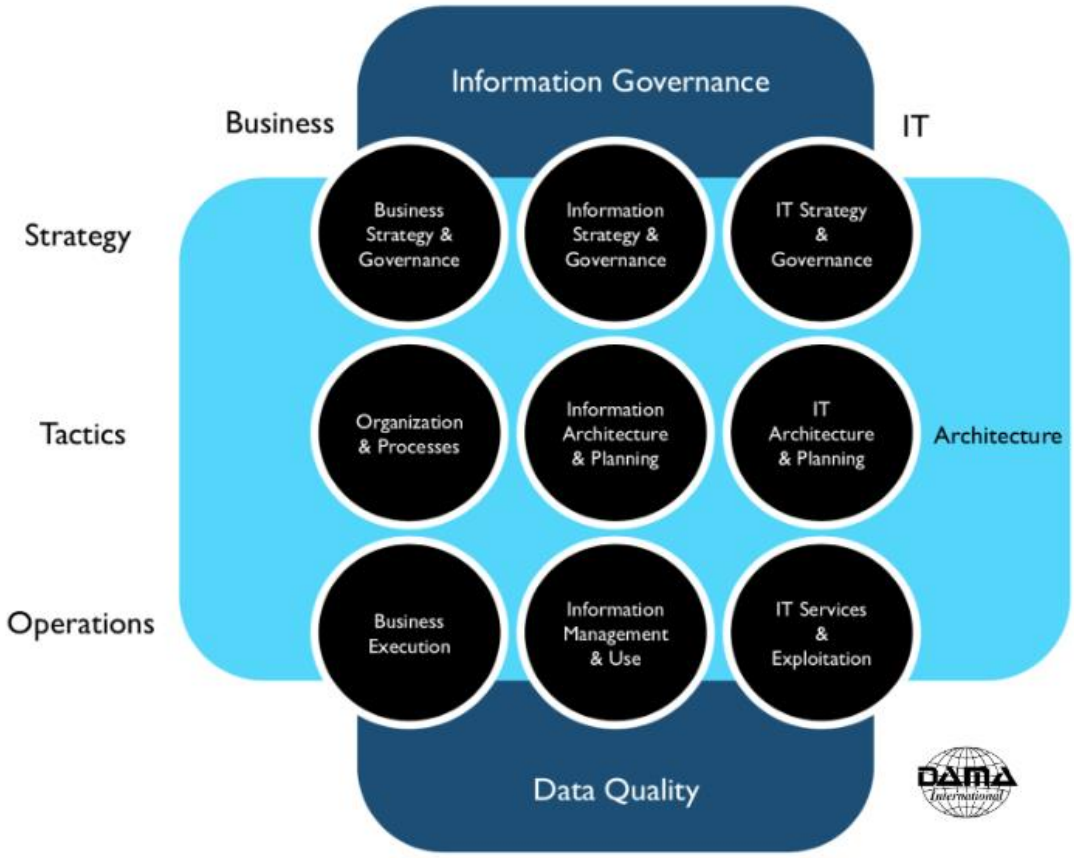


Figure 4 Amsterdam Information Model

3.3 The DAMA-DMBOK Framework

Consists of three visuals:

- **The DAMA Wheel:** Governance in the centre for consistency within and balance between the Knowledge Areas

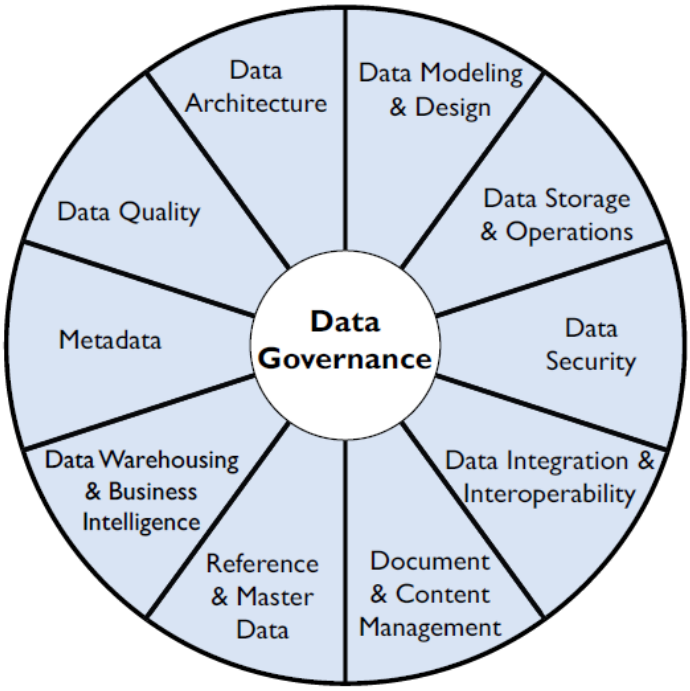


Figure 5 The DAMA-DMBOK2 Data Management Framework (The DAMA Wheel)

- **The Environmental Factors Hexagon:** Relationships between people, process and technology. Goals and principles in the centre.

Figure 6 DAMA Environmental Factors Hexagon

DMBOK Version 1 explains:

The Basic Environmental Elements are:

1. *Goals and Principles: The directional business goals of each function and the fundamental principles that guide performance of each function.*
2. *Activities: Each function is composed of lower level activities. Some activities are grouped into sub-activities. Activities are further decomposed into tasks and steps.*
3. *Deliverables: The information and physical databases and documents created as interim and final outputs of each function. Some deliverables are essential, some are generally recommended, and others are optional depending on circumstances.*
4. *Roles and Responsibilities: The business and IT roles involved in performing and supervising the function, and the specific responsibilities of each role in that function. Many roles will participate in multiple functions.*

The supporting Environmental Elements are:

5. *Techniques: Common and popular methods and procedures used to perform the processes and produce the deliverables. Practices and Techniques may also include common conventions, best practice recommendations, and alternative approaches without elaboration.*
6. *Tools: Categories of supporting technology (primarily software tools), standards and protocols, product selection criteria and common learning curves.*
7. *Organization and Culture: These issues might include:*
    - *Management Metrics measures of size, effort, time, cost, quality, effectiveness, productivity, success, and business value.*
    - *Critical Success Factors.*
    - *Reporting Structures.*
    - *Contracting Strategies.*
    - *Budgeting and Related Resource Allocation Issues.*
    - *Teamwork and Group Dynamics.*
    - *Authority and Empowerment.*
    - *Shared Values and Beliefs.*
    - *Expectations and Attitudes.*
    - *Personal Style and Preference Differences.*
    - *Cultural Rites, Rituals and Symbols.*
    - *Organizational Heritage.*
    - *Change Management Recommendations*.

Each Knowledge Area Context Diagram was made up using the following matrix:

Data Management Functions	Goals and Principles	Activities	Primary Deliverables	Roles and Responsibilities	Technology	Practices and Techniques	Organization and Culture
Data Governance							
Data Architecture Management							
Data Development							
Data Operations Management							
Data Security Management							
Reference and Master Data Management							
Data Warehousing and Business Intelligence Management							
Document and Content Management							
Meta-data Management							
Data Quality Management							

The DMBOK Version 2 generic Knowledge area context diagram:

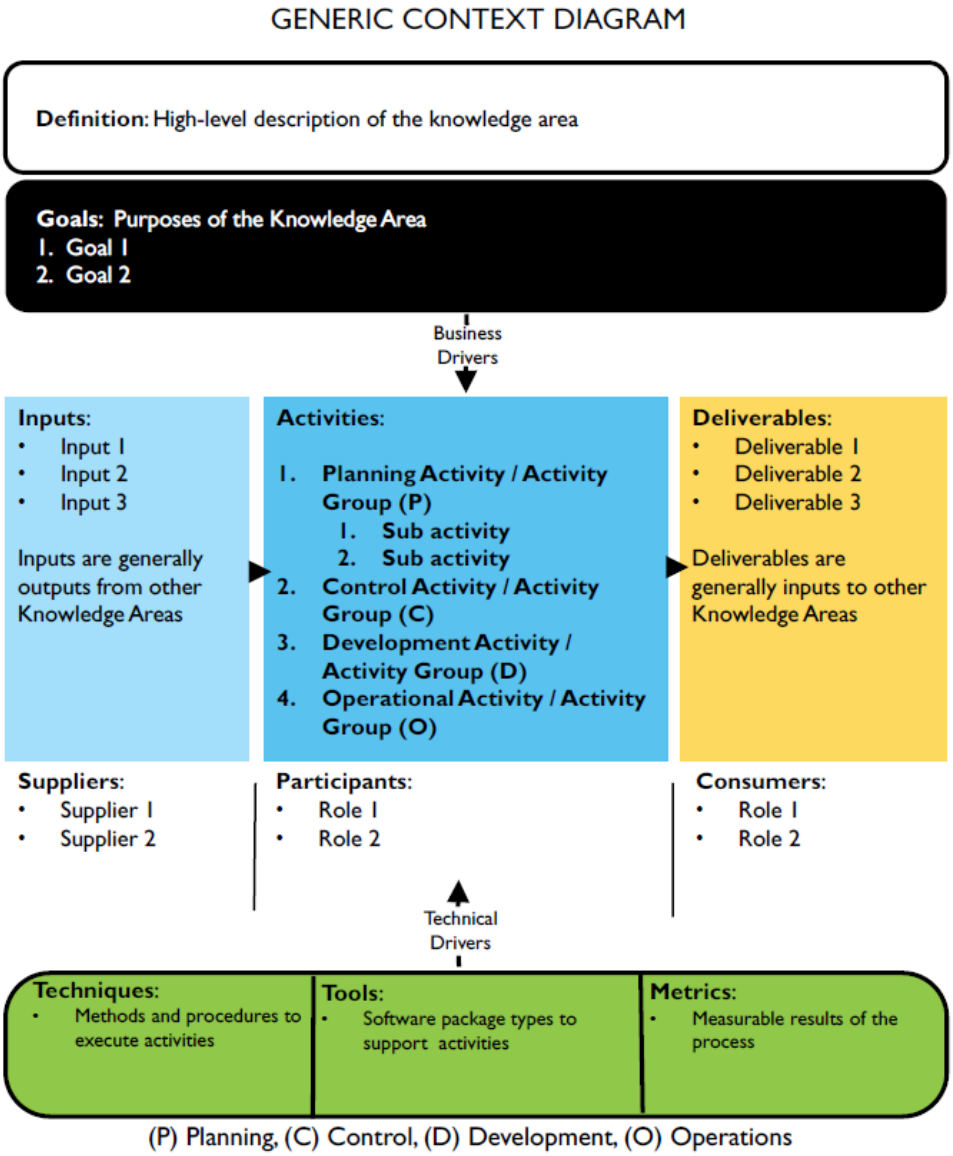


Figure 7 Knowledge Area Context Diagram

- Definition
- Goals
- Activities that drive goals classified in 4 phases:
  - (P) Planning – Strategic and tactical
  - (C) Control Activities – ensure quality of data, integrity, reliability, security of systems
  - (D) Development activities – organised around system lifecycle (SDLC)
  - (O) Operational Activities – support systems and processes through which data is accessed and used
- Inputs
- Deliverables
- Roles and Responsibilities (SFIA)
- Suppliers
- Consumers

- Participants
- Tools
- Techniques
- Metrics

3.4 DMBOK Pyramid (Aiken)

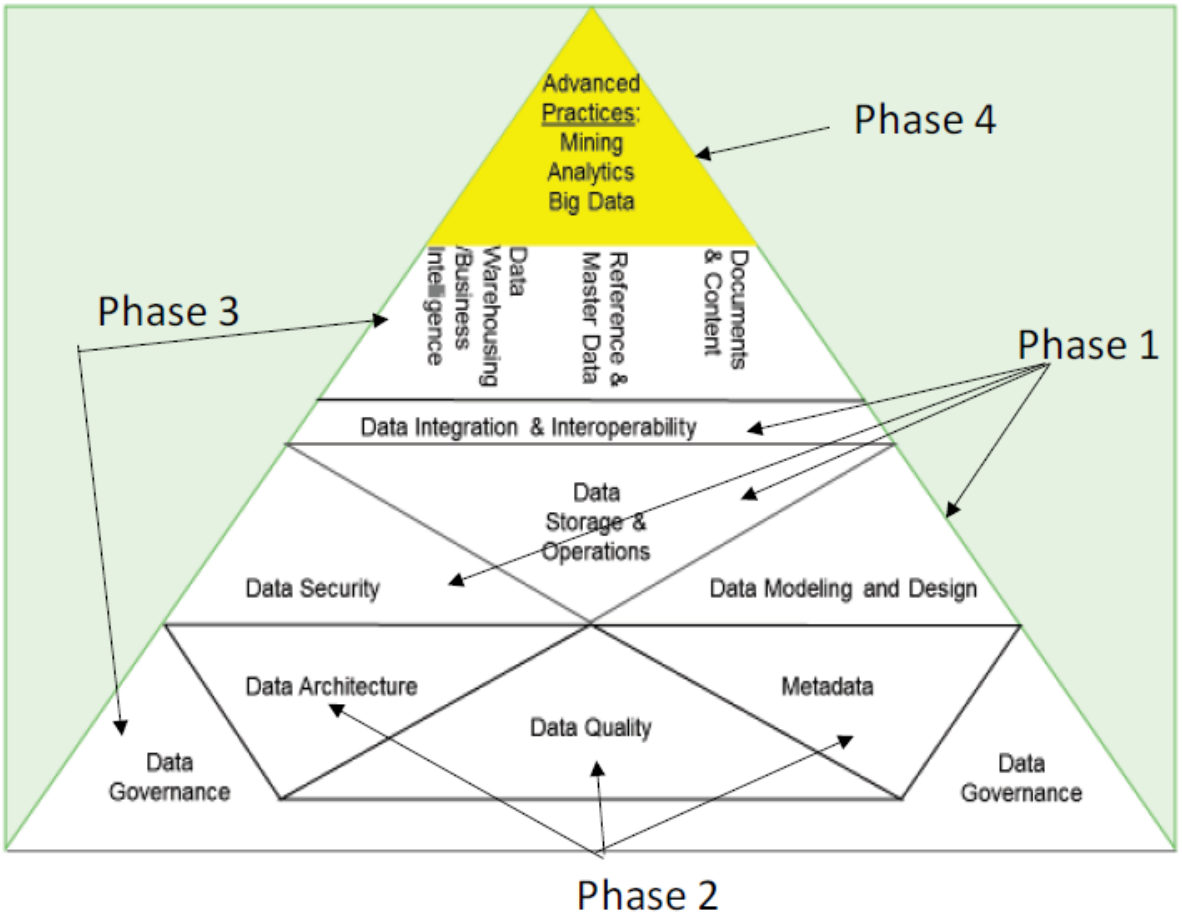


Figure 8 Purchased or Built Database Capability

Describes how an organisation progresses to Data Management maturity:

3.5 DAMA Data Management Framework Evolved

Sue Geuens explored the dependencies between the knowledge areas, and developed Figure 9, DAMA Functional Area Dependencies.

- BI/Analytics depend on all other functions.
- DW and Master data depend on feeding systems/applications.
- Depend on reliable data quality, design and integration.
- Governance is the foundation on which all functions are dependent.

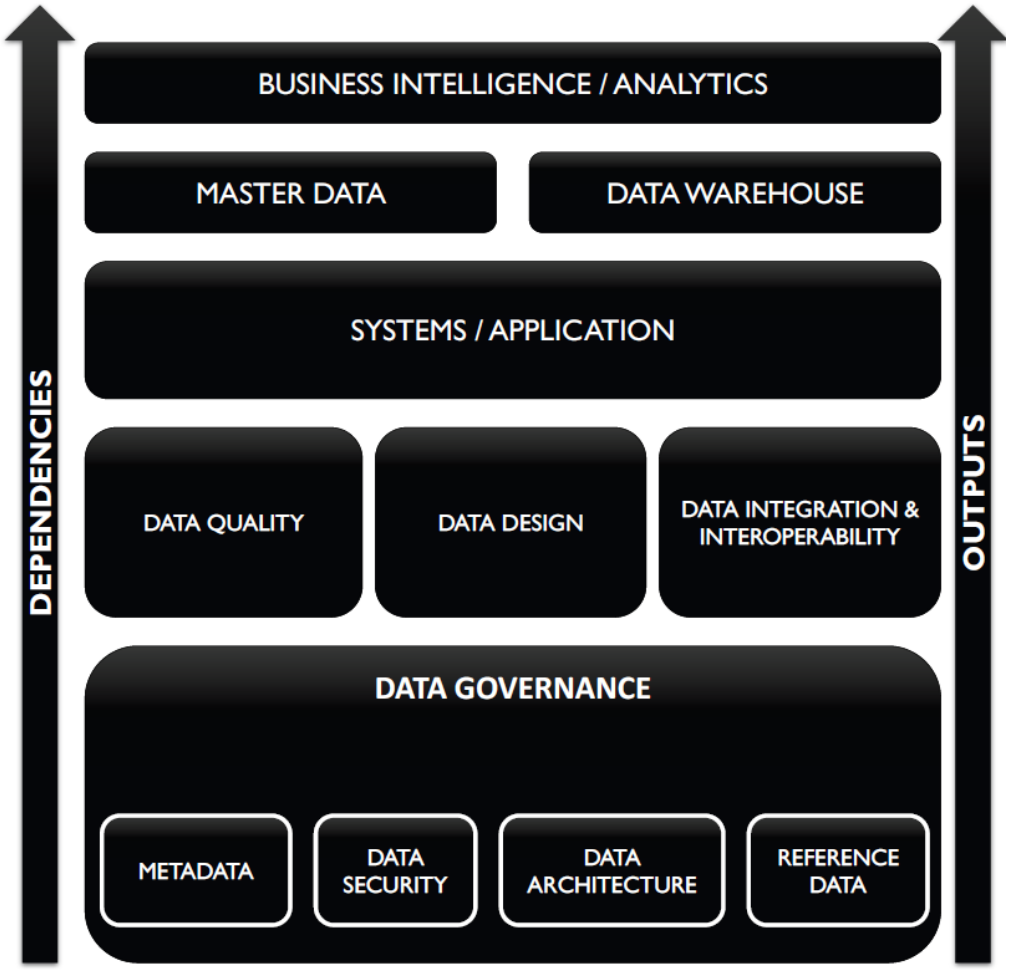


Figure 9 DAMA Functional Area Dependencies

DAMA Data Management Function Framework (DMBOK V2 Figure 10 below):

- Guiding purpose of Data Management – Deriving value from data assets
- Deriving value requires lifecycle management – Depicted in the centre of the diagram
- Foundational activities support the lifecycle
- A data governance program provides oversight and containment, and enables the organisation to be data driven, as it puts in place the strategy and supporting principles, policies, procedures and stewardship practices.

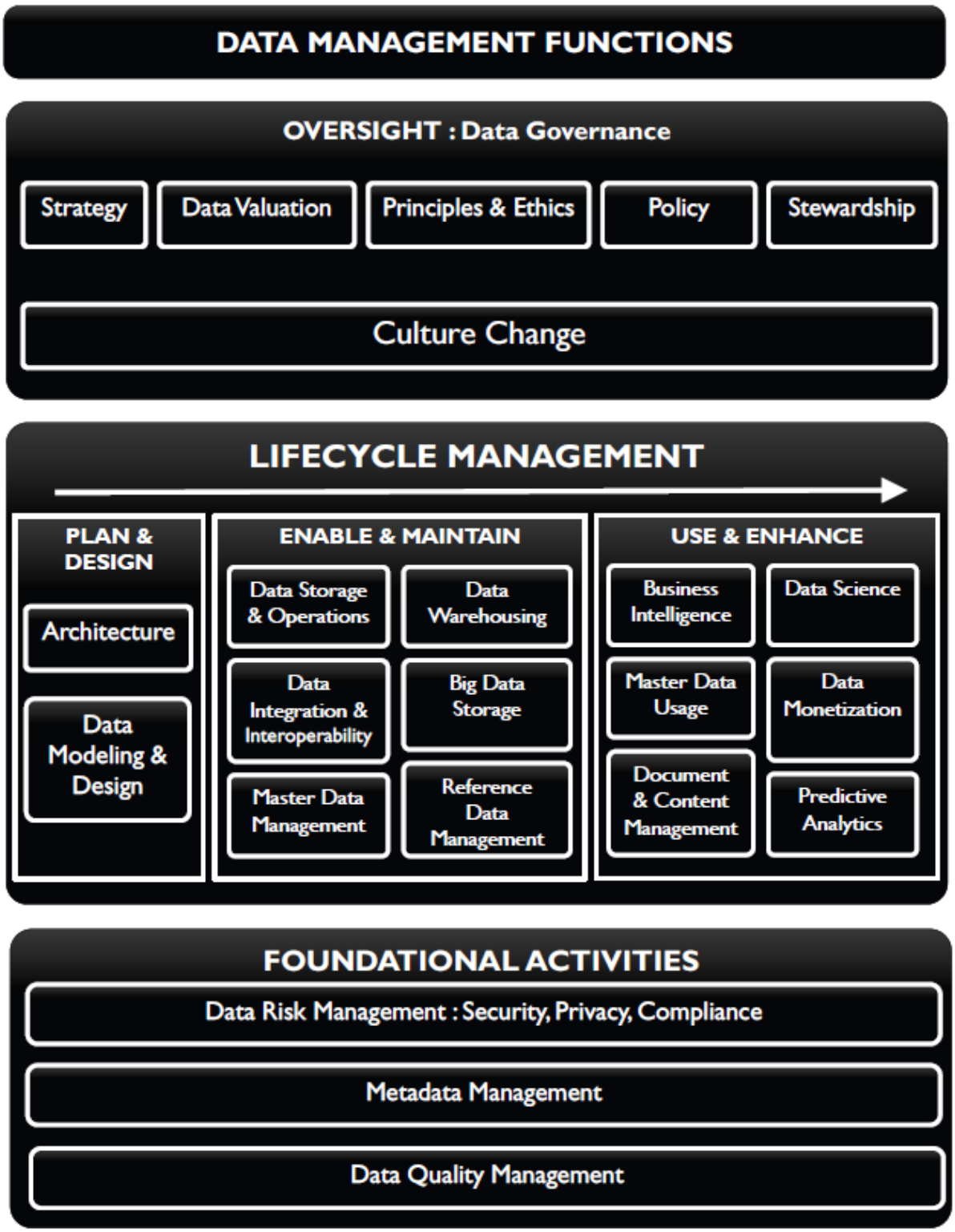


Figure 10 Data Management Function Framework

(Note: There are errors is this figure in the DMBOK 2. Corrected here and in the Revised Version.)

Evolution of the DAMA Wheel:

Core activities surrounded by lifecycle and usage activities, within the strictures of governance:

- Core activities in centre

- Surrounded by lifecycle and usage activities
- Lifecycle management: Plan and Enable
- Uses come from lifecycle activities
- Data Governance provides oversight and containment

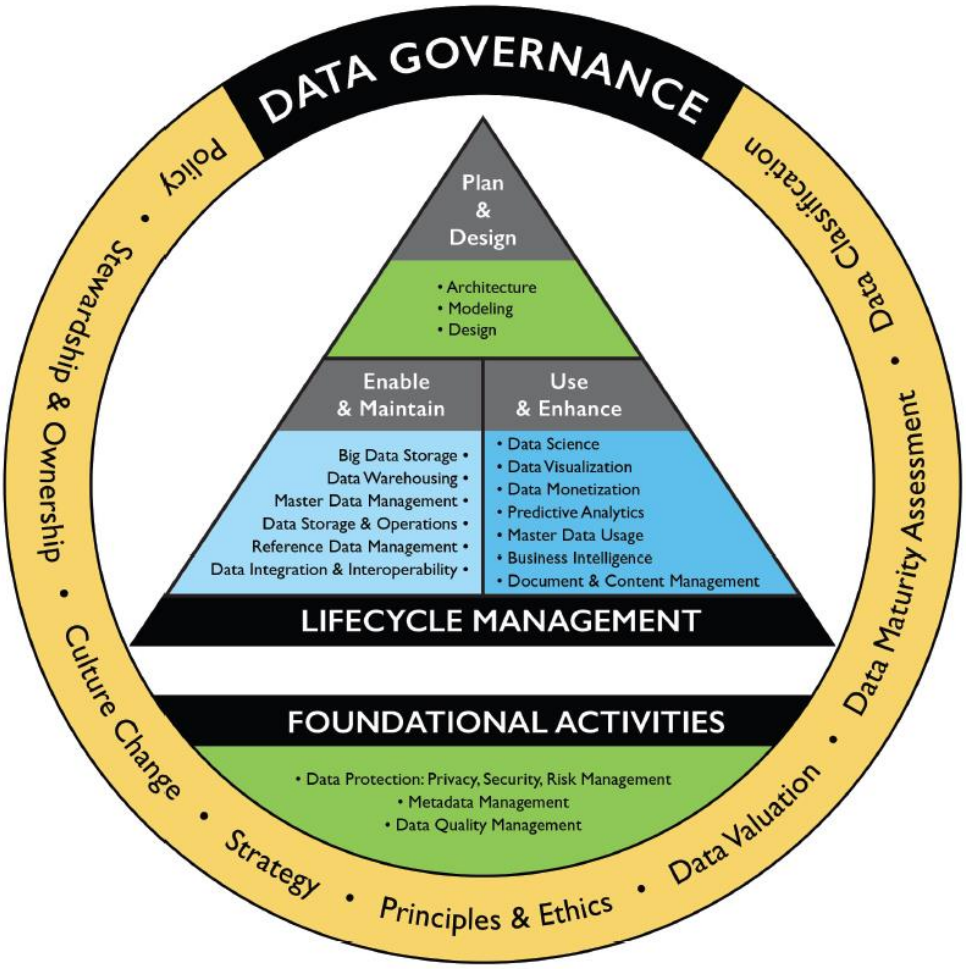


Figure 11 DAMA Wheel Evolved

DM Functions and Data Lifecycle Management

Relationships with additional content of the knowledge areas. Data management enables organisations to get value from their data. This require data lifecycle management, and these activities are in the centre of the diagram.

4 DAMA and the DMBOK

DAMA was founded to address the need for reliable data management practices.

The DMBOK supports DAMA's mission by:

- Providing a functional framework for the implementation of enterprise data management practices
- Enabling a common vocabulary for data management practices
- Serving as a fundamental guide for the CDMP exams

Knowledge Areas describe the scope and context of data management activities. They intersect with each other as data moves horizontally within an organisation.

1. **Data Governance** provides direction and oversight for data management by establishing a system of decision rights over data that accounts for the needs of the enterprise. (Ch 3 – **11%**)
2. **Data Architecture** defines the blueprint for managing data assets by aligning with organisational strategy to establish strategic data requirements and designs to meet these requirements. (Ch 4 – **6%**)
3. **Data Modelling and Design** is the process of discovering, analysing, representing and communicating data requirements in a precise form called a Data Model. (Ch 5 – **11%**)
4. **Data Storage and Operations** includes the design, implementation and support of stored data to maximise its value. Operations provide support throughout the data lifecycle from planning to disposal of data. (Ch 6 – **6%**)
5. **Data Security** ensures that data privacy and confidentiality are maintained, that data is not breached, and that data is accessed appropriately. (Ch 7 – **6%**)
6. **Data Integration and Interoperability** includes processes related to the movement and consolidation of data within and between data stores, applications and organisations. (Ch 8 – **6%**)
7. **Document and Content Management** includes planning, implementation and control activities used to manage the lifecycle of data and information found in a range of unstructured media, especially documents needed to support legal and regulatory compliance requirements. (Ch 9 – **6%**)
8. **Reference and Master Data** includes ongoing reconciliation and maintenance of core shared data to enable consistent use across systems of the most accurate, timely and relevant version of the truth about essential business entities. (Ch 10 – **10%**)
9. **Data Warehousing and Business Intelligence** includes the planning, implementation and control processes to manage decision support data to enable knowledge workers to get value from data via analysis and reporting. (Ch 11 – **10%**)
10. **Metadata** includes planning, implementation and control activities to enable access to high quality, integrated Metadata, including definitions, models, data flows and other information critical to understanding data and the systems through which it is created, maintained and accessed. (Ch 12 – **11%**)
11. **Data Quality** includes planning and implementation of quality management techniques to measure, assess and improve the fitness of data for use within the organisation. (Ch 13 – **11%**)
12. **Data Handling Ethics** describes the central role that data ethics plays in making informed, socially responsible decisions about data and its uses. Awareness of the ethics of data collection, analysis and use should guide all data management professionals. (Ch 2 – **2%**)
13. **Big Data and Data Science** describes the technologies and business processes that emerge as our ability to collect and analyse large and diverse data sets increases. (Ch 14 – **2%**)
14. **Data Management Maturity Assessment** outlines an approach to evaluating and improving an organisation's data management capabilities. (Ch 15)
15. **Data Management Organisation and Role Expectations** provide best practices and considerations for organising data management teams and enabling successful data management practices. (Ch 16)

16. **Data Management and Organisational Change Management** describes how to plan for and successfully move through the cultural changes that are necessary to embed effective data management practices within and organisation. (Ch 17)

Chapter 2
Data Handling Ethics

1 Introduction

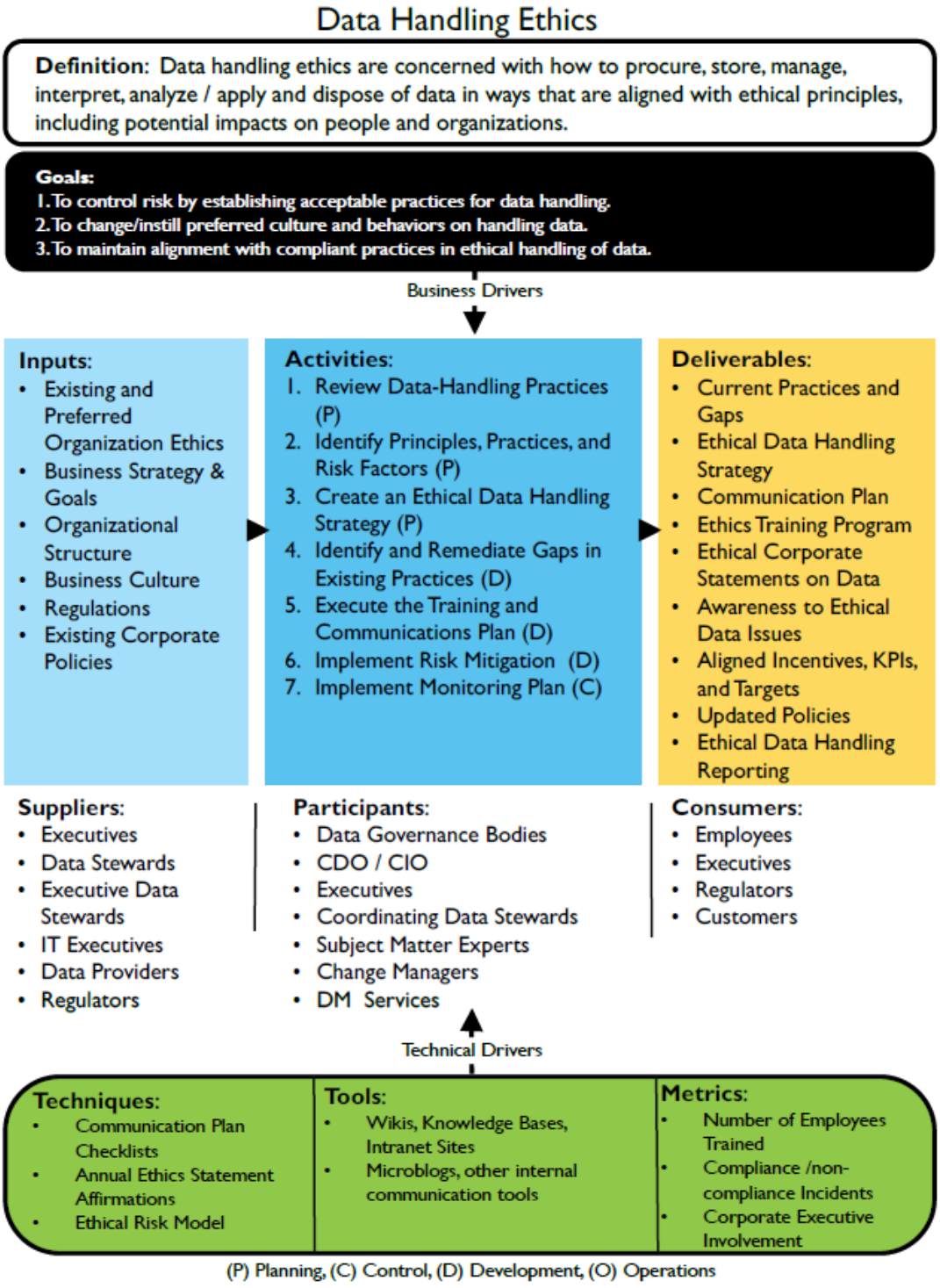


Figure 12 Context Diagram: Data Handling Ethics

Ethics of data handling are complex. Core Concepts:

- **Impact on people:**
  - Imperative to maintain the quality and reliability of individuals' data
  - Used to make decisions that impact people's lives
- **Potential for misuse:**
  - Negative effects on people and organisations
- **Economic value of data:**
  - Determine how and by whom that value of data can be accessed

2 Business Drivers

Ethics means "doing it right when no one is looking".

Ethical handling of data is a competitive business advantage for an organisation:

- Increased trustworthiness and improved relationships – a competitive advantage
- Better relationships with stakeholders
- Reduced risk of misuse or a criminal breach
- Responsibilities when sharing data
- Organisation wide commitment to handling data ethically

Chief Data Officer, Chief Risk Officer, Chief Privacy Officer, Chief Analytics Officer focus on controlling risk by establishing acceptable practices for data handling.

3 Essential Concepts

3.1 Ethical Principles for Data

- **Respect for persons**: Treat people in a way that respects their dignity and autonomy as human individuals. Protect the dignity of those with "diminished Autonomy"
- **Beneficence**: Firstly, do no harm. Secondly, maximise possible benefits and minimise possible harms
- **No maleficence**: Minimise harm
- **Justice**: The fair and equitable treatment of people

European Data Protection Supervisor, 2015 – the "engineering, philosophical, legal and moral implications" of developments in **Data Processing and Big Data**:

- Future-oriented regulation of data processing and respect for the rights to privacy and data protection
- Accountable controllers who determine personal information processing
- Privacy conscious engineering and design of data processing products and devices
- Empowered individuals

3.2 Principles behind Data Privacy Law

Privacy regulations have been driven by human rights violations experienced during WW2 in Europe. The European Union's Organisation for Economic Co-operation and Development (OECD) established 8 Guidelines and Principles for Fair Information Processing.

- Limitations on data collection
- An expectation that data will be of a high quality
- The requirement that when data is collected it will be for a specific purpose
- Limitations on data usage
- Security safeguards
- An expectation of openness and transparency
- The right of an individual to challenge the accuracy of data related to himself or herself
- Accountability of organisations to follow the guidelines

These underpin the GDPR:

3.2.1 General Data Protection Regulation of the EU

GDPR Principle	Description of Principle
Fairness, Lawfulness, Transparency	Personal data shall be processed lawfully, fairly, and in a transparent manner in relation to the data subject.
Purpose Limitation	Personal data must be collected for specified, explicit, and legitimate purposes, and not processed in a manner that is incompatible with those purposes.
Data Minimization	Personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed.
Accuracy	Personal data must be accurate, and where necessary, kept up-to-date. Every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purpose for which they are processed, are erased or rectified without delay.
Storage Limitation	Data must be kept in a form that permits identification of data subjects [individuals] for no longer than is necessary for the purposes for which the personal data are processed.
Integrity and Confidentiality	Data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures.
Accountability	Data Controllers shall be responsible for, and be able to demonstrate compliance with [these principles].

Table 1 GDPR Principles

Other focusses for fair information practices include:

- Simplified consumer choice to reduce burden placed on consumers
- Recommendation to maintain comprehensive data management practices throughout the lifecycle
- A Do Not Track option
- Requirements for affirmative express consent
- Concerns regarding the data collection capabilities of large platform providers; transparency and clear privacy notices and policies
- Individuals' access to data
- Educating consumers about data privacy practices

- Privacy by Design

3.3 Online Data in an Ethical Context

- **Ownership of data:** Right to control own personal data
- **The right to be forgotten:** To have information about an individual erased from the web, particularly to adjust online reputation
- **Identity:** Identity must be correct, or one can opt for a private identity
- **Freedom of speech online:** Expression rather than online bullying or insulting

3.4 Risks of Unethical Data Handling Practices

Data should be measured against Data Quality dimensions such as accuracy and timeliness to ensure it is trustworthy. Unethical data practices include:

- **Timing**: Omission or inclusion of data points based on timing, e.g., market manipulation of end-of-day stock price.
- **Misleading Visualisations:**
  - e.g., Changing scale, comparing facts without clarifying their relationship
  - Unclear definitions or Invalid Comparison
  - Provide context which informs meaning.
  - Statistical "smoothing" or over-training of a statistical model.
- **Bias:** an inclination of outlook:
  - Data collection for pre-defined results.
  - Biased use of collected data
  - Hunch and search
  - Biased sampling methodology
  - Context and Culture
- **Unclear Definitions or Invalid Comparisons:**
  - Business definitions and context of comparisons is necessary
- **Transforming and Integrating Data:**
  - Limited knowledge of data's origin and lineage
  - Data of poor quality
  - Unreliable metadata
  - No documentation of data remediation history
- **Obfuscation / Redaction of Data:** Risk of exposure is common in the following:
  - Data aggregation
  - Data marking
  - Data masking

3.5 Establishing an Ethical Data Culture

- **Review current state data handling practices:**
  - Understand degree of current practices to ethical and compliance drivers
  - Employees understanding of ethical implications of existing practices
- **Identify principles, practices and risk factors**
- **Create an ethical data handling strategy and roadmap:** Ethical principles and expected behaviour related to data
  - **Values statements:** What the organisation believes in
  - **Ethical data handling principles:** How the organisation approaches challenges related to data
  - **Compliance framework:** Factors that drive organisational compliance obligations

- **Risk assessment:** Likeliness of a specific problem occurring
- **Training and communications:** Train on code of ethics. Communications must reach all employees
- **Roadmap:** Timeline of activities that can be approved by management
- **Approach to auditing and monitoring**

- **Adopt a socially responsible ethical risk model:** BI, analytics and data science responsible for data that describes:
  - Who people are (race, country or origin, religion)
  - What people do (Political, social, criminal activities)
  - How people live (Money, purchases, who they communicate with)
  - How people are treated (outcomes of analysis which privilege or prejudice them for future business)
- **The risk model is used to determine**:
  - Whether to execute the project
  - How to execute the project
  - Actively identify potential risks
  - Protect whistle-blowers
  - Identify possible bias

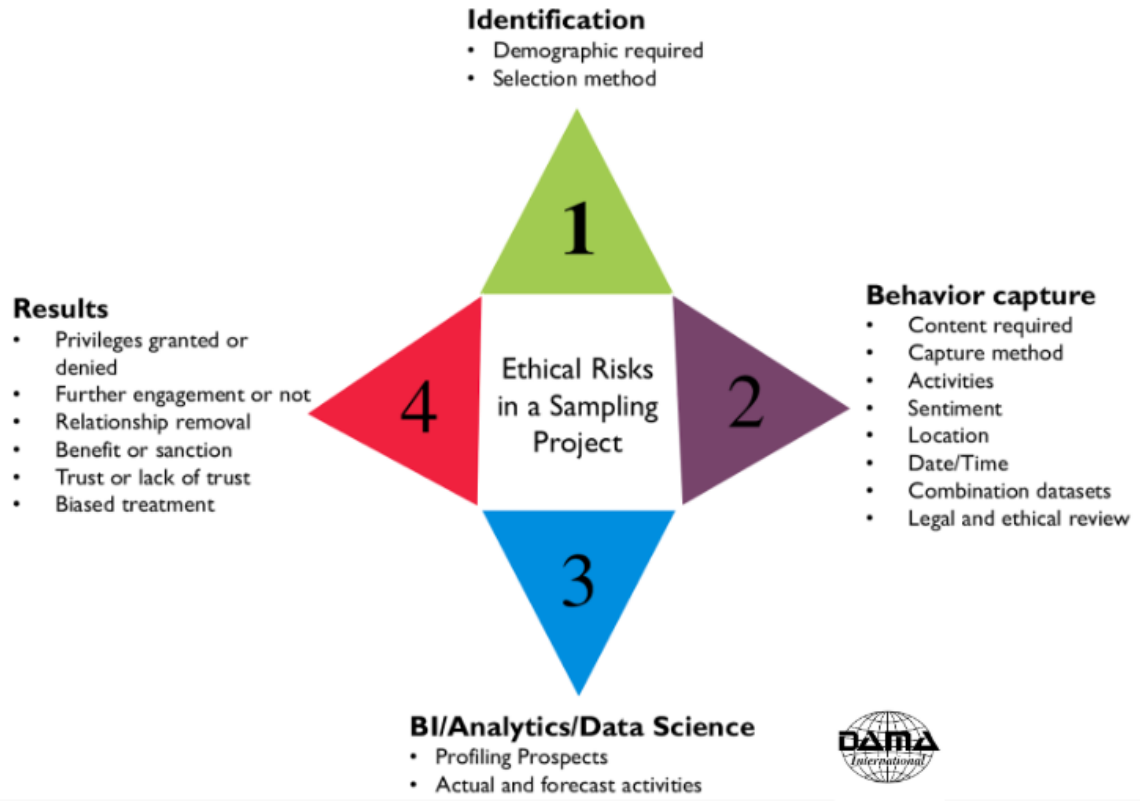


Figure 13 Ethical Risk Model for Sampling Projects

Address all potential ethical risks in the areas of consideration, with a particular focus on negative effects on customers or citizens.

3.6 Data Ethics and Governance

Oversight falls under Data Governance and Legal counsel:

- Keep up-to-date on legal changes

- Ensure employees are aware of their obligations
- Data Governance must set standards and policies to provide oversight of data handling practices
- Particularly to review plans and decisions proposed by BI, analytics and Data Science
- CDMP Certification requires subscription to code of ethics

Chapter 3
Data Governance

1 Introduction

The purpose of Data Governance is to ensure that data is managed properly, according to policies and best practices:

- **Strategy:** Defining, communicating and driving execution of Data Strategy and Data Governance
- **Policy:** Setting and enforcing policies relating to data and Metadata lifecycle management (creation, access, usage, security, quality, documentation and purging)
- **Standards:** Setting and enforcing Data Management standards
- **Oversight:** Providing hands-on observation, audit and correction in key data management areas like quality, privacy, security, and data management policies and procedures
- **Compliance:** Ensuring the organisation can meet the requirements of regulations
- **Issue management:** Identifying, defining, escalating and resolving issues related to data security, data access, data quality, regulatory compliance, data ownership, policy, standards, terminology or data governance procedures
- **Data management projects:** Sponsoring efforts to improve data management practices
- **Data asset valuation:** Setting standards and processes to define business value of assets

The Data Governance function develops policies and procedures, cultivates stewardship practices and organisational change management.

Organisational culture must be changed to value data and data management activities. Formal change management is required.

1.1 Business Drivers

Reducing Risk:

- **General risk management:** Oversight of the risks data poses to finances or reputation, including response to legal and regulatory issues
- **Data security:** Protection of data assets through controls
- **Privacy:** Control of private, confidential and Personal Identifying Information (PII) through policy and compliance monitoring

Improving Processes:

- **Regulatory compliance:** The ability to respond efficiently and consistently
- **Data Quality improvement:** Improved business performance by making data more reliable
- **Metadata Management:** Business glossary and other Metadata
- **Efficiency in development projects:** SDLC improvements
- **Vendor management:** Contracts dealing with data such as cloud, data purchase, sales of data as a product, and outsourcing

The Data Governance function is separate from the IT governance function, and is ongoing.

- IT makes decisions about hardware, software and the technical architecture

- IT aligns IT investments and strategies with enterprise goals and strategies
- COBIT framework (Control Objectives for Information and Related Technologies) for IT governance

- Figure 14 Context Diagram: Data Governance and Stewardship

1.2 Goals and Principles

Goals:
1. Enable an organization to manage its data as an asset.
2. Define, approve, communicate, and implement principles, policies, procedures, metrics, tools, and responsibilities for data management.
3. Monitor and guide policy compliance, data usage, and management activities.

To achieve these goals a Data Governance function must be:

- **Sustainable:** Ongoing, and depends on business leadership, sponsorship and ownership. DG is an enterprise function that requires organisational commitment
- **Embedded:** DG activities incorporated into development methods, use of analytics, management of Master Data and risk management
- **Measured:** to show positive financial impact

Foundational principles**:**

- **Leadership and strategy:** Committed leadership, driven by enterprise business strategy
- **Business-driven:** Data Governance is a business function and must govern IT decisions relating to data
- **Shared responsibility:** Across all Data Management Knowledge Areas
- **Multi-layered:** All levels from local to enterprise
- **Framework-based:** Establish an operating framework
- **Principle-based:** Principles are the basis of policies. Principles can mitigate resistance

Data Governance provides oversight and containment

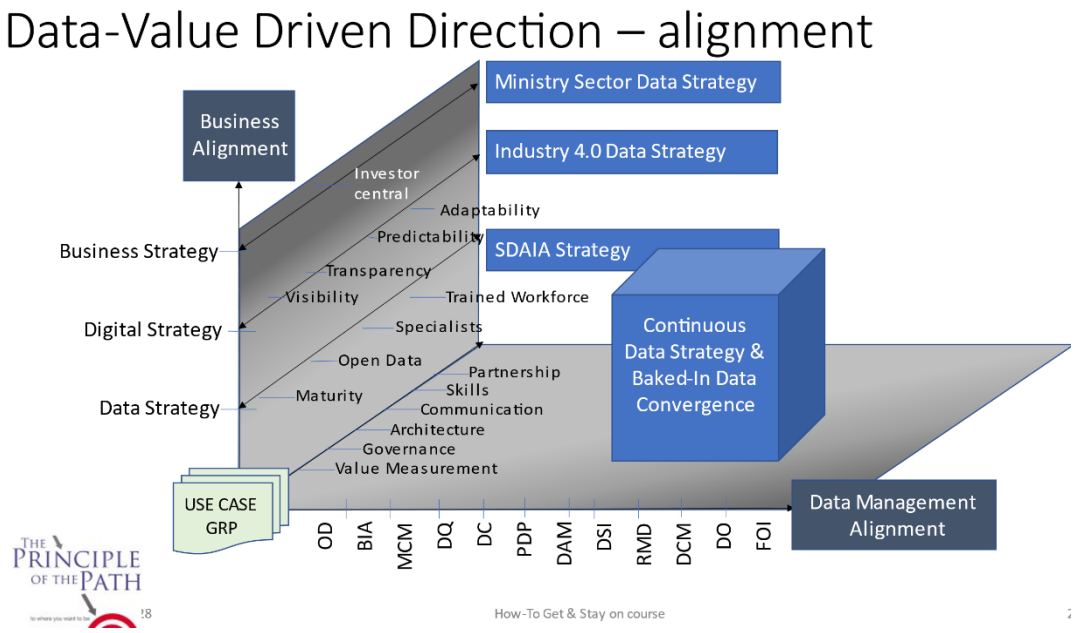
DMBOK Quote
- Data governance activities provide oversight and containment, through strategy, principles, policy, and stewardship.
- They enable consistency through data classification and data valuation

Why is this important from a Governance point of view?
- To ensure that Data remains a strategic imperative for the business
- The Data Risks are identified & managed

How will Data Governance do this?
- Implement a Data Governance function and supporting deliverables

2020/05/28 Modelware Systems & DAMA SA 3

From How-to Govern Data Course by Howard Diesel



The Data-Value Driven Direction (DVD Way) By Howard Diesel

1.3 Essential concepts

Ensuring data is managed without directly executing data management. **Separation of duties between oversight and execution.**

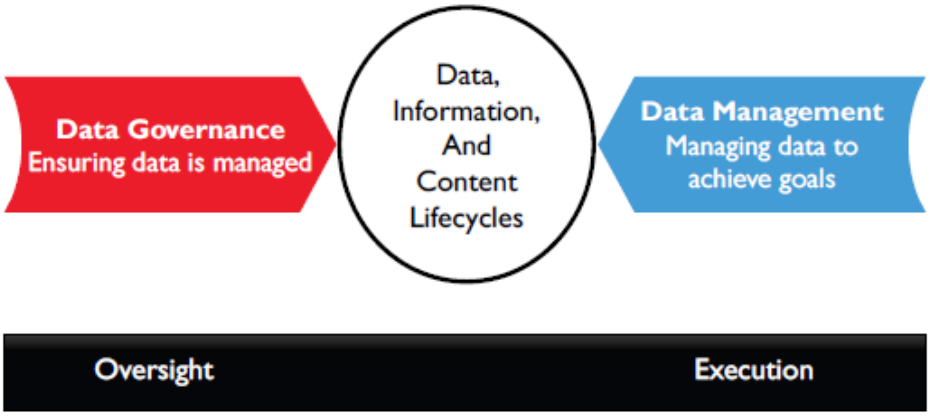


Figure 15 Data Governance and Data Management

1.3.1 Data-centric Organisation

A data-centric organisation values data as an asset and manages data through all phases of its lifecycle, including project development and ongoing operations.

Principles:

- Manage data as a corporate asset
- Incentivise Data management best practices across the organisation
- Enterprise data strategy is aligned to overall business strategy
- Continually improve data management processes

1.3.2 Data Governance Organisation

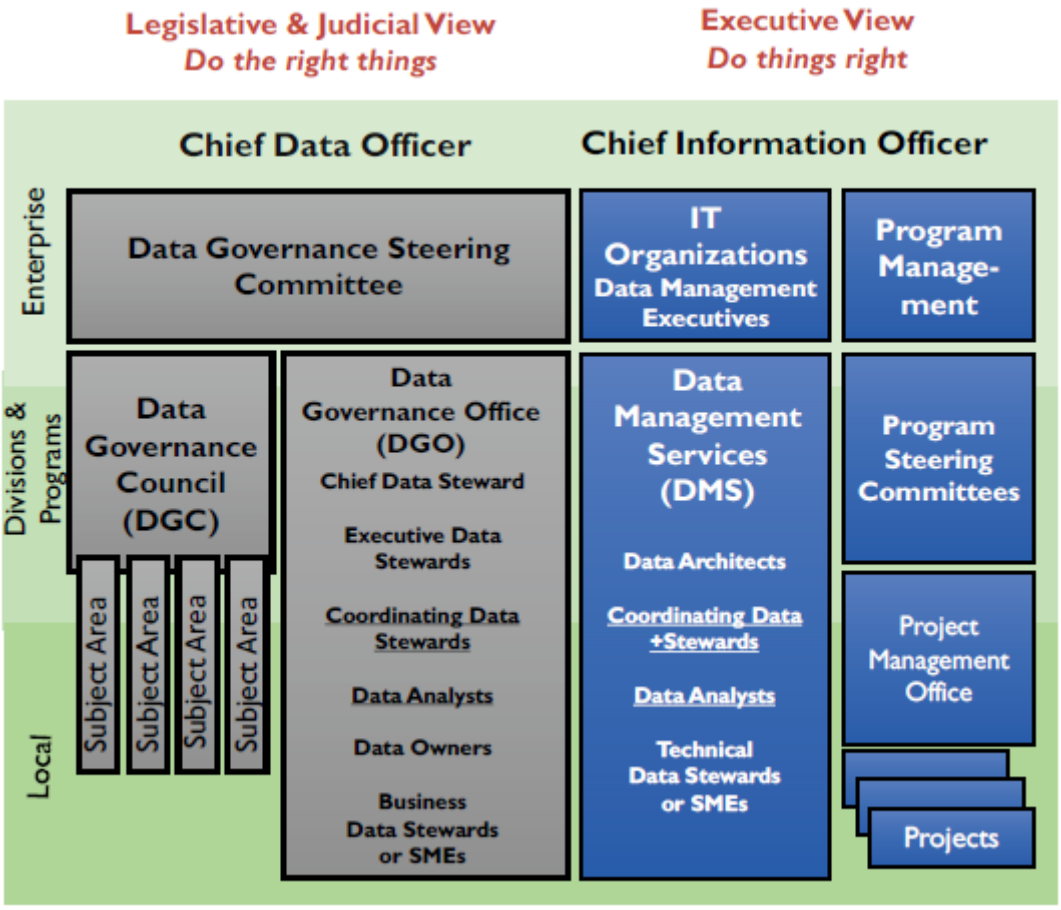
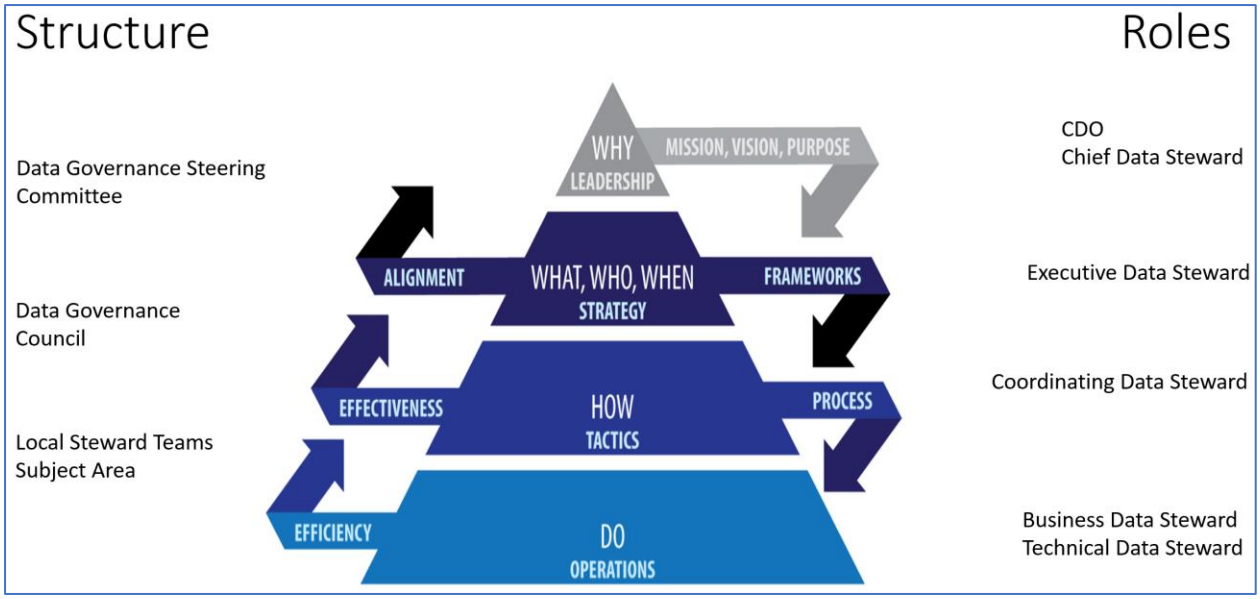


Figure 16 Data Governance Organisation Parts

Data Governance Body	Description
Data Governance Steering Committee	The primary and highest authority organization for Data Governance in an organization, responsible for oversight, support, and funding of Data Governance activities. Consists of a cross-functional group of senior executives. Typically releases funding for Data Governance and Data Governance-sponsored activities as recommended by the DGC and CDO. This committee may, in turn, have oversight from higher-level funding or initiative-based steering committees.
Data Governance Council (DGC)	Manages Data Governance initiatives (e.g., development of policies or metrics), issues, and escalations. Consists of executives from either the organization (in a centralized or federated model) or corresponding business unit (in a replicated model). See Figure 17.
Data Governance Office (DGO)	Ongoing focus on enterprise-level data definitions and data management standards across all DAMA-DMBOK Knowledge Areas. Consists of coordinating roles that are labeled as *data stewards* or *custodians*, and *data owners*.
Data Stewardship Teams	Communities of interest focused on one or more specific subject-areas or projects, collaborating or consulting with project teams on data definitions and data management standards related to the focus. Consists of business and technical data stewards and data analysts.
Local Data Governance Committee	Large organizations may have divisional or departmental Data Governance councils working under the auspices of an Enterprise DGC. Smaller organizations should try to avoid such complexity.

Table 4 Typical Data Governance Committees / Bodies



Alignment of Structure and Roles with Leadership, Strategic, Tactical and Operational activities (Author of Graphic Unknown)

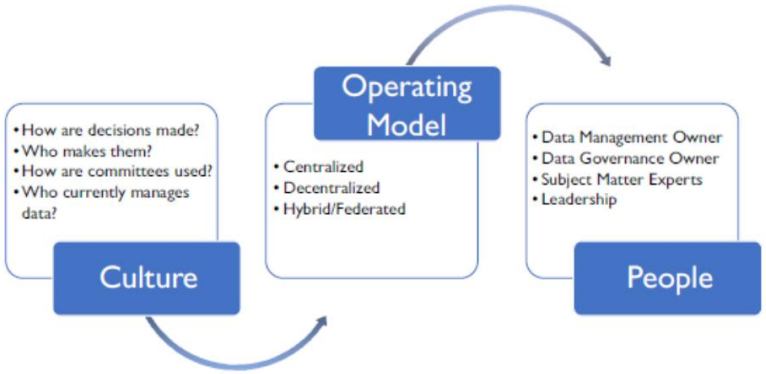


Figure 106 Assess Current State to Create an Operating Model (From DMBOK V2 Chapter 16)

1.3.3 Data Governance Operating Model Types

(More detail in Chapter 16)

- **Centralised**: One Central EIM / DG Team with Councils within the Business
- **Decentralised/Replicated**: One DG Team per Business Organization Structure
- **Hybrid**: One Central EIM and By Business Organization Structure (Chapter 16)
- **Federated**: Unite in a federation – individual autonomy. Centralised strategy, functions in the BUs
- **Self-organising/Network**: Non-Invasive Model based on RACI (Chapter 16)

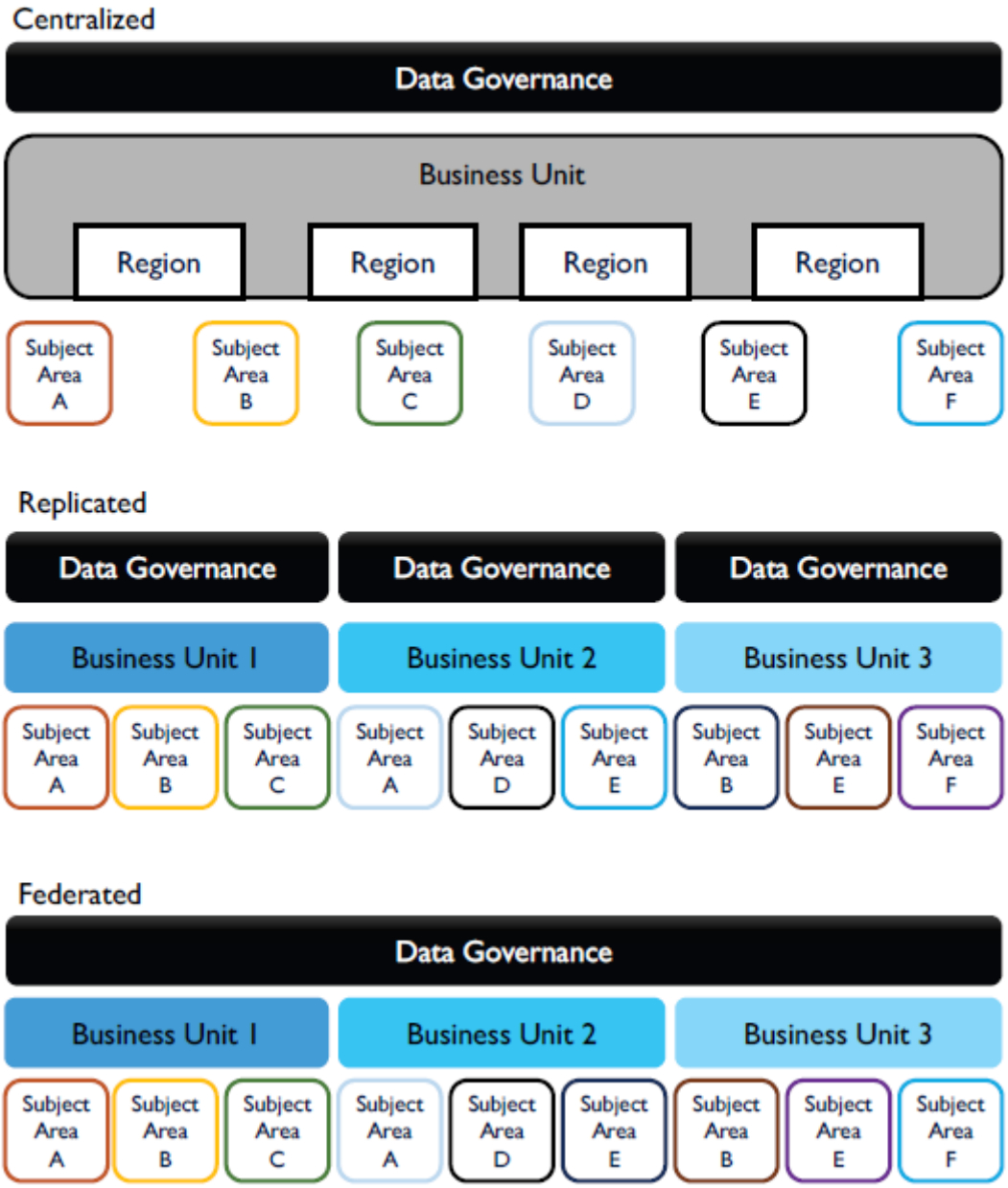


Figure 17 Enterprise Data Governance Operating Framework Examples

1.3.4 Data Stewardship

Data Stewardship describes accountability and responsibility for data and processes that ensure effective control and use of data assets. Owner, custodian or trustee.

Activities include:

- Creating and managing core Metadata
- Documenting rules and standards
- Managing data quality issues
- Executing operational data governance activities

1.3.5 Types of Data Stewards

A steward manages the property of another person. Data stewards manage the data assets on behalf of others and in the best interests of the organisation. (McGilvray, 2008). Data Stewards are

accountable and responsible for data governance and have a portion of their time dedicated to these activities.

Types of Data Stewards:

- **Chief Data Stewards**: Chair DG bodies, act as CDO, be Executive Sponsors
- **Executive Data Stewards**: Senior managers who serve on the Data Governance Council
- **Enterprise Data Stewards**: Oversight of a data domain across business functions
- **Business Data Stewards**: Business professionals, subject matter experts
- **A Data Owner**: A business person who is accountable for decisions about data within their domain
- **Technical Data Stewards**: IT professionals operating in one of the knowledge areas e.g., DBAs, BI Specialists, Data Integration specialists, Data Quality analysts, Metadata administrators
- **Coordinating Data Stewards**: Lead teams of business and technical data stewards in discussions across teams and with Executive Data Stewards

Data Principles: Data Stewards: (Extra, not in the DMBOK, source unknown)

- ***Enterprise data must be modelled:***
  - *Modelled, named and defined according to standards across all business divisions*
  - *Effort made by management to share data – not maintain redundant data*
  - *Originating data stewards recognise needs of downstream processes*
- ***Enterprise data must be maintained close to the source:***
  - *Create and maintain as close to the source as possible*
  - *Data Quality standards applied to approved reliability levels as defined by business units*
- ***Enterprise data must be safe and secured:***
  - *In all electronic formats must be safeguarded on requirements and compliance guidelines*
  - *Guidelines are determined by business stewards of Enterprise data*
  - *Backup and recovery measures*
- ***Enterprise data must be accessible**:*
  - *Enterprise data and metadata shall be readily available and accessible to all except where restricted*
  - *Business stewards of enterprise data are responsible for defining the types of individuals, and levels of access privileges*
  - *Information Security will be responsible for the implementation of proper security controls*
- ***Metadata will be recorded and utilised:***
  - *All projects utilise defined Metadata for data naming, data modelling and database design.*
  - *Data Governance is responsible for the Metadata program*

1.3.6 Data Policies

- Data Policies are directives that codify principles and management intent into fundamental rules governing the creation, acquisition, integrity, security, quality and use of data and information
- Global

- Describe the "What" of data Governance (Standards and procedures describe "How")
- Should be few data polices, and should be stated briefly and directly

1.3.7 Data Asset Valuation

The process of understanding and calculating the economic value of an intangible asset like data to the organisation.

Data sets are not interchangeable (fungible) between organisations. How an organisation gets value from customer data can be a competitive differentiator.

The following phases of the data lifecycle (acquiring, storing, administering and disposing) involve costs. Data only brings value when it is used but incurs risk management costs.

Ways to measure value:

- **Replacement cost** of data lost in a breach or disaster
- **Market value:** As a business asset at the time of a merger or acquisition
- **Identified opportunities**: The income to be gained
- **Selling data** as a package, or the insights to be gained
- **Risk cost**: Potential penalties, remediation costs and litigation expenses

Generally Accepted Information Principles (GAIP) translated from Generally Accepted Accounting Principles (GAAP).

Principle	Description
Accountability Principle	An organization must identify individuals who are ultimately accountable for data and content of all types.
Asset Principle	Data and content of all types are assets and have characteristics of other assets. They should be managed, secured, and accounted for as other material or financial assets.
Audit Principle	The accuracy of data and content is subject to periodic audit by an independent body.
Due Diligence Principle	If a risk is known, it must be reported. If a risk is possible, it must be confirmed. Data risks include risks related to poor data management practices.
Going Concern Principle	Data and content are critical to successful, ongoing business operations and management (i.e., they are not viewed as temporary means to achieve results or merely as a business by-product).
Level of Valuation Principle	Value the data as an asset at a level that makes the most sense or is the easiest to measure.
Liability Principle	There is a financial liability connected to data or content based on regulatory and ethical misuse or mismanagement.
Quality Principle	The meaning, accuracy, and lifecycle of data and content can affect the financial status of the organization.
Risk Principle	There is risk associated with data and content. This risk must be formally recognized, either as a liability or through incurring costs to manage and reduce the inherent risk.
Value Principle	There is value in data and content, based on the ways these are used to meet an organization's objectives, their intrinsic marketability, and/or their contribution to the organization's goodwill (balance sheet) valuation. The value of information reflects its contribution to the organization offset by the cost of maintenance and movement.

Table 5 Principles for Data Asset Accounting

2 Data Governance Activities

2.1 Define Data Governance for the Organisation

- DG must support business strategy and goals
- Enterprise Data Strategy is informed by business strategy and goals
- DG enables shared responsibility for data decisions
- Clear understanding of what and who are governed and who is governing
- DG is most effective when it is an enterprise effort

2.2 Perform Readiness Assessment

- **Data management maturity**: Understand what the organisation does with data, measure current management capabilities and capacity
- **Capacity to change**: Measure capacity of organisation to change, and identify possible resistance points
- **Collaborative readiness**: measures collaboration across functional areas
- **Business alignment**: how well data use aligns with business strategy

2.3 Perform Discovery and Business Alignment

- **Discovery Activity**
    - Identify and assess effectiveness of existing policies and guidelines (Risks, behaviours, implementation)
    - Identify opportunities for DG to improve usefulness of data
- **Business Alignment** attaches business benefits to Data Governance function elements
- **Data Quality analysis**
    - Identify issues and risks associated with poor quality data
    - Identify business processes at risk from poor quality data
    - Financial and other benefits from creating a Data Quality program as part of DG efforts
- Assessment of **data management practices**
- Derive a list of **DG requirements** which will drive **DG strategy** and tactics

2.4 Develop Organisational Touch Points

Touch points that support alignment of an enterprise data governance and data management outside the direct authority of the Data Governance function. This supports the organisation's cohesiveness in managing data, increasing its nimbleness to use data.

- **Procurement and contracts**: Enforce standard contract language for data management contracts
- **Budget and funding**: Prevent duplicate acquisitions and ensure optimisation of data assets
- **Regulatory compliance**: The Data Governance function understands and works within regulatory requirements. Requires ongoing monitoring
- **SDLC/development framework**: Data Governance function identifies control points where enterprise policies, processes and standards can be developed in system or development lifecycles

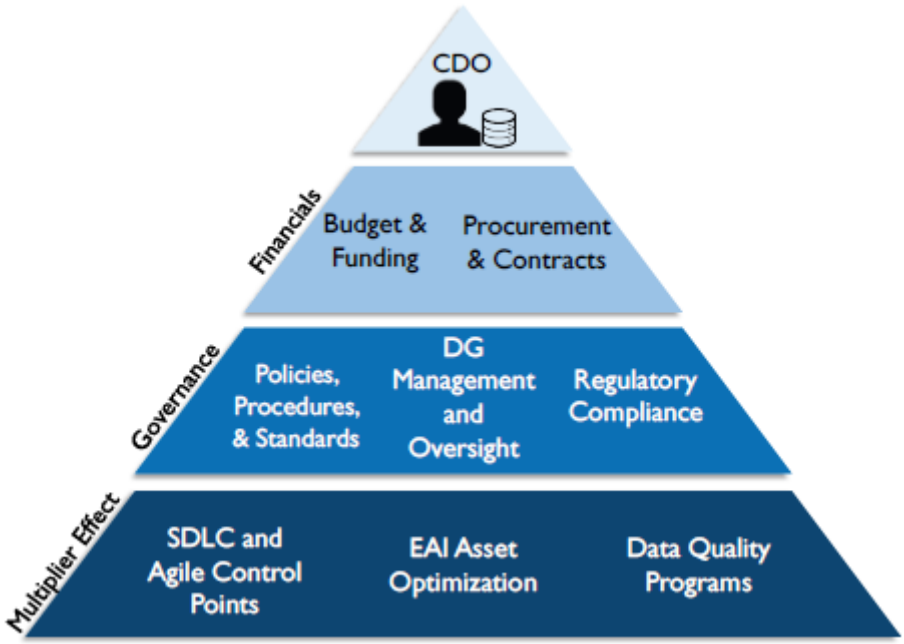


Figure 18 CDO Organisational Touch Points

2.5 Develop Data Governance Strategy

Defines the scope and approach to governance efforts. Deliverables:

- **Charter:** Business drivers, vision, mission and principles of data governance
- **Operating framework and accountabilities:** Structure and responsibilities
- **Implementation roadmap:** Timeframes
- **Plan for operational success:** Describe target state

2.6 Define the DG Operating Framework

Consider the following areas when constructing the organisations operating model:

- **Value of data to the organisation:** Important if the organisation sells data
- **Business model:** Decentralised, centralised, international
- **Cultural factors:** Acceptance of discipline and adaptability to change
- **Impact of regulation:** Level of regulation of the organisation

Layers of governance: determine where accountability resides for stewardship activities and data ownership.

The DG Operating framework defines:

- The interaction between Governance Organisation and people responsible for data management initiatives
- Engagement of change management initiatives to introduce DG
- Model for issue resolution pathways through governance

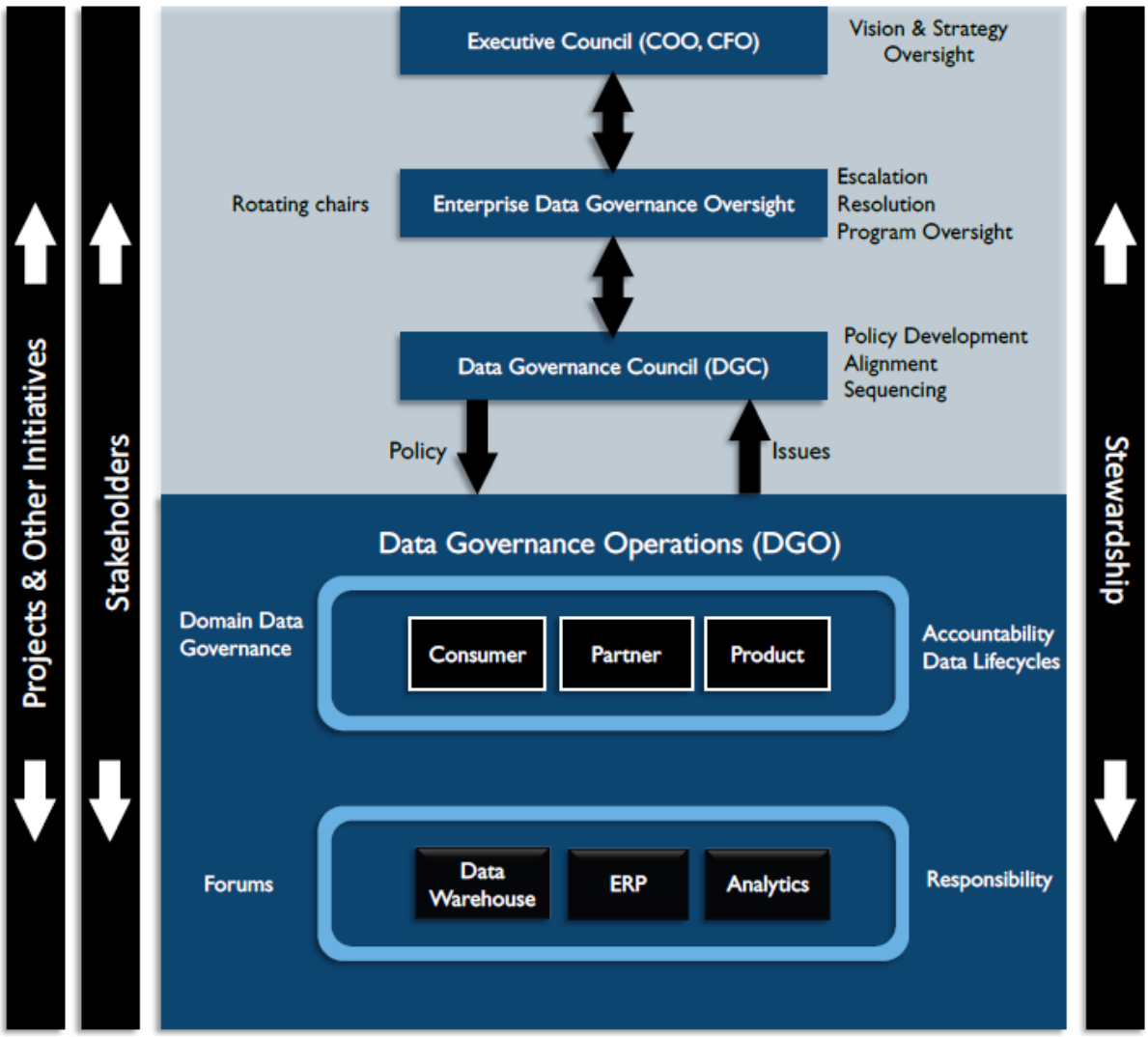


Figure 19 An Example of an Operating Framework

2.7 Develop Goals, Principles and Policies

Derived from the DG Strategy to the desired future state. They are:

- Drafted by data management professionals and/or business policy staff
- Refined by data stewards and management
- Data Governance Council conducts final review, revision and adoption

Data policies must be effectively communicated, monitored, enforced and periodically re-evaluated. Data Governance Council may delegate this authority to the Data Stewardship Steering Committee.

2.8 Underwrite Data Management Projects

Promote enterprise wide data management improvements by articulating the ways they improve efficiency and reduce risk. Should be priority for organisations wanting more value from their data.

The DGC helps define the business case and oversees data management improvement project status and progress in coordination with the Project Management Office (PMO).

Data Management projects are part of the IT Project portfolio.

The DGC coordinates projects with enterprise wide scope such as Master Data Management, Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM).

Data Management requirements must be captured in the planning and design phases of the SDLC

2.9 Engage Change Management

Organisational Change Management (OCM) is required to bring about change in the organisation's systems and processes. A mature organisation in Change Management builds clear vision, leads and monitors from the top, and designs and manages small changes with feedback. Involves collaboration in whole organisation.

- **Planning:**
  - Stakeholder analysis
  - Gain sponsorship
  - Communications approach to resistance to change
- **Training:** Creating and executing training plans for Data Governance functions
- **Influencing systems development:** Engage with PMO to add DG steps to the SDLC
- **Policy implementation:** Communicate policies and the organisation's commitment to Data Management activities
- **Communications**:
  - Promoting the value of data assets
  - Monitoring and acting on feedback about DG activities
  - Implementing data management training
  - Measuring the effects of change management in **5 key areas**:
    - Awareness of the need to change
    - Desire to participate and support the change
    - Knowledge about how to change
    - ability to implement new skills and behaviours
    - Reinforcement to keep change in place
- **Implementing new metrics and KPIs:** Employee incentives

2.10 Engage in Issue Management

The process of identifying, quantifying, prioritising and resolving DG issues:

- **Authority:** Questions around rights and procedures
- **Change management escalations:** Arising from the change management process
- **Compliance:** Issues needing compliance requirements
- **Conflicts:** Policies, procedures, business rules, names, definitions, stakeholder interests etc.
- **Conformance:** Issues related to conformance to policies, standards, architecture etc.
- **Contracts:** Negotiation and review
- **Data security and identity**: Privacy and confidentiality issues and breaches
- **Data quality:** Detection and resolution of DQ issues

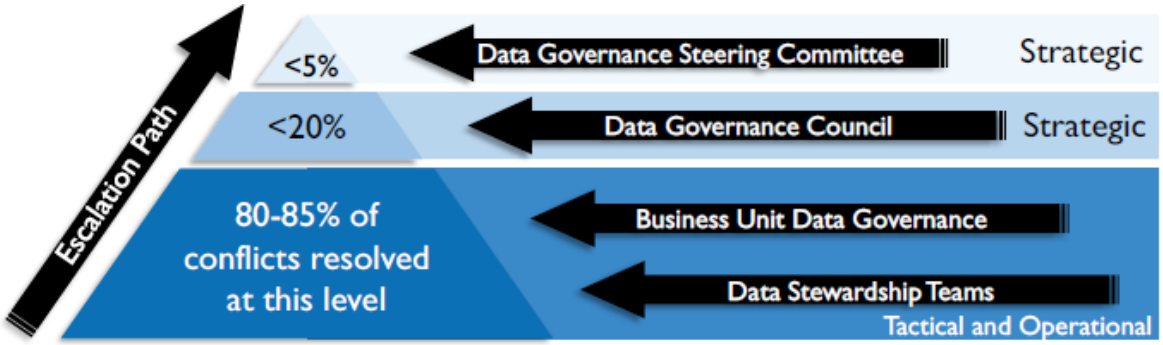


Figure 20 Data Issue Escalation Path

Develop control mechanisms and procedures for:

- Identifying, capturing, logging and updating issues
- Assignment and tracking of action items
- Documenting stakeholder viewpoints and resolution alternatives
- Determining, documenting and communicating issue resolutions
- Facilitating objective, neutral discussions where all viewpoints are heard
- Escalating issues to higher levels of authority

2.11 Assess Regulatory Compliance Requirements

Part of the DG Function is to monitor and ensure regulatory compliance. DG guides the implementation of controls to monitor and document compliance with data-related regulations.

Examples of global regulations which impact data management practices.

- Accounting standards
- BCBS 239 (Basel Committee on Banking Supervision) and Basel II – for banks
- CPG 235 (Australian Prudential Regulation Authority APRA) – banks and insurance
- PCI-DSS – The Payment Card Industry Data Security Standards
- Solvency II - European Union Rules for insurance, similar to Basel II
- Privacy Laws

Evaluate the implications of the regulations

- Relevance of regulation to the organisation
- What constitutes compliance, and what policies and procedures are required?
- When is compliance achieved, and how and when is it monitored?
- Can industry standards achieve compliance?
- How is compliance demonstrated?
- Risks and penalties for non-compliance
- How is non-compliance identified and reported?
- How is non-compliance managed and rectified?

2.12 Implement Data Governance

There are many complex activities which need to be coordinated using a roadmap with timeframes. Schedules may differ in a federated model with different business units based on differing levels of engagement, maturity and funding.

Prioritised DG activities in the early stage:

- Define activities for high-priority goals
- Business glossary, document terminology and standards
- Coordinate with Enterprise and Data Architecture to understand the data and systems
- Assign financial value to data assets to enable better decision making and understanding

2.13 Sponsor Data Standards and Procedures

Standard: Something set up and established by authority as a rule for measure of quantity, weight, extent, value or quality.

Standards help define DQ by providing a means of comparison.

Standards have the potential to simplify processes, as a decision is made once, and codified in a set of assertions (the standard).

Enforce standards to promote consistent results from the processes using them. Data can be measured against standards. The DGC or Data standards steering Committee should audit DM activities as part of the SDLC approval process or by schedule.

Standards difficulties in organisations: DG Standards should be mandatory.

- Politicised, causing goals to be lost
- Organisation not practiced at developing or enforcing DG standards
- Value of implementing standards is not recognised
- No knowledge how to implement standards

Different forms of data standards:

- How a field should be populated
- Rules governing relationships between fields
- Acceptable and unacceptable values
- Format

Process:

- Standards are drafted by data management professionals
- Reviewed, approved and adopted by the DGC or a Data Standards Steering Committee (A delegated workgroup)
- Document with capturing organisational knowledge in mind.

Data standards must be communicated, monitored, reviewed and updated. There must be a means to enforce them.

The DGC or DSSC should audit DM activities for standards compliance on a defined schedule or as part of the SDLC approval processes.

Data Management Procedures: The documented methods, techniques and steps followed to accomplish specific activities that produce certain outcomes and supporting artifacts.

Concepts can be standardised within all the Data Management Knowledge Areas.

Examples of concepts to be standardised in the KAs:

- **Data Architecture:** Enterprise data models, tools, system naming conventions

- **Data Modelling & Design:** Data model management procedures, naming conventions, definition standards, standard domains, and standard abbreviations
- **Data Storage & Ops:** Tool standards, database recovery, business continuity, database performance, external data acquisition
- **DII:** Standard methods and tools
- **Doc & Content:** Standard procedures, enterprise taxonomies, legal discovery, retention periods, electronic signatures, report distribution approaches
- **Ref & Master Data:** RDM control procedures, systems of record, use mandates, standards for entity resolution
- **DW&BI:** Tool, processing standards and procedures, report / visualisation formatting standards, Big Data handling standards
- **Metadata:** Standards business and technical metadata to be captured, integration
- **Data Quality:** DQ rules, standard measuring methodologies, data remediation standards and procedures
- **Big Data & Data Science:** Data source identification, authority, acquisition, system of record, sharing and refresh

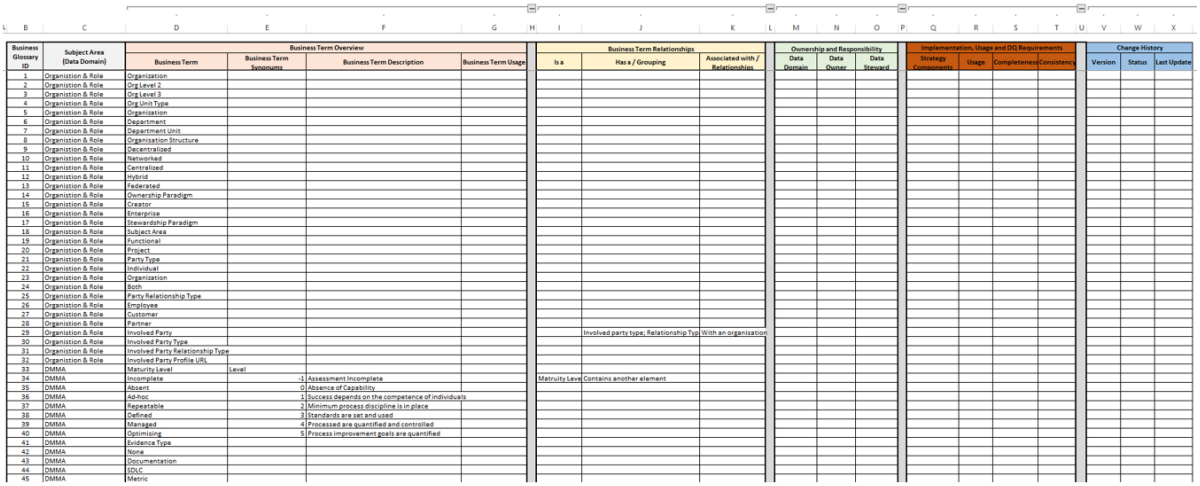
2.14 Develop a Business Glossary

A **Business Glossary** is a list of terms, definitions and other Metadata such as synonyms, metrics, lineage, security and privacy classification, business rules, the responsible steward for that term etc. Business Data Stewards are generally responsible for business glossary content.

Objectives of business glossary:

- Enable common understanding of core business concepts and terminology
- Reduce the risk that data will be misused due to inconsistent understanding of the business concepts
- Improve alignment between technology assets and the business organisation
- Maximise search capability and enable access to documented institutional knowledge

Example of a Business Glossary



Business Glossary ID	Subject Area (Data Domain)	Business Term Overview: Business Term	Business Term Synonums	Business Term Description	Business Term Usage	Business Term Relationships: Is a	Has a / Grouping	Associated with / Relationships	Ownership and Responsibility: Data Domain	Data Owner	Data Steward	Implementation, Usage and DQ Requirements: Strategy Components	Usage	Completeness	Consistency	Change History: Version	Status	Last Update
1	Organization & Role	Organization																
2	Organization & Role	Org Level 2																
3	Organization & Role	Org Level 3																
4	Organization & Role	Org Unit Type																
5	Organization & Role	Organization																
6	Organization & Role	Department																
7	Organization & Role	Department Unit																
8	Organization & Role	Organization Structure																
9	Organization & Role	Decentralized																
10	Organization & Role	Networked																
11	Organization & Role	Centralized																
12	Organization & Role	Hybrid																
13	Organization & Role	Federated																
14	Organization & Role	Ownership Paradigm																
15	Organization & Role	Creator																
16	Organization & Role	Enterprise																
17	Organization & Role	Stewardship Paradigm																
18	Organization & Role	Subject Area																
19	Organization & Role	Functional																
20	Organization & Role	Project																
21	Organization & Role	Party Type																
22	Organization & Role	Individual																
23	Organization & Role	Organization																
24	Organization & Role	Both																
25	Organization & Role	Party Relationship Type																
26	Organization & Role	Employee																
27	Organization & Role	Customer																
28	Organization & Role	Partner																
29	Organization & Role	Involved Party						Involved party type; Relationship Type	With an organization									
30	Organization & Role	Involved Party Type																
31	Organization & Role	Involved Party Relationship Type																
32	Organization & Role	Involved Party Profile URL																
33	DMMA	Maturity Level	Level															
34	DMMA	Incomplete		-1 Assessment Incomplete														
35	DMMA	Absent		0 Absence of Capability														
36	DMMA	Ad-hoc		1 Success depends on the competence of individuals														
37	DMMA	Repeatable		2 Minimum process discipline is in place														
38	DMMA	Defined		3 Standards are set and used														
39	DMMA	Managed		4 Processed are quantified and controlled														
40	DMMA	Optimising		5 Process improvement goals are quantified														
41	DMMA	Evidence Type																
42	DMMA	None																
43	DMMA	Documentation																
44	DMMA	SDLC																
45	DMMA	Metric																

Business Glossary Excel Template from the How-to Govern Data Course by Howard Diesel

2.15 Coordinate with Architecture Groups

The Data Governance Council is a stakeholder (possibly sponsors and approves) for Data Architecture artefacts such as a business-oriented enterprise data model.

Enterprise Data model is sponsored, reviewed and approved by DGC.

Enterprise Data Architecture Steering Committee or Architecture Review Board (ARB) appointed by DGC to oversee DA Program and its iterative projects.

- Enterprise Data model developed by Data Architects and Data Stewards in subject area teams
- Changes or extensions to EDM proposed and developed by Data Steward Teams
- EDM must align with business strategy and data strategy

The Enterprise data model must be reviewed, approved and formally adopted by the DGC. It must align to key business strategies, processes and systems. Doing things right.

2.16 Define Data Asset Valuation Method

DGC organises and standardises the effort to put monetary value to data.

- Information gaps represent business liabilities
- Business value of the missing data can be the cost of closing the gaps
- Develop models to estimate value of information that does not exist
- Build value estimates into the data strategy road map
    - Justifies business cases for root cause DQ solutions
    - Business cases for other DG Initiatives

2.17 Embed Data Governance

The organisation accepts the governance of data, and the processes and funding are in place to enable the continued performance of the DG framework.

- **Goal of the DGO:**
    - Embed behaviours related to managing data as an asset in processes
    - Operations plan to implement and operate DG activities
    - Activities, timing and techniques to ensure success
- **Sustainability of the DG organisational framework:**
    - Processes and funding in place
    - Organisation accepts the governance of data
    - DG is measured, monitored and obstacles overcome
- **Create a Data Governance Community of Interest**
    - Deepens the understanding of DG
    - Helpful in early years of governance

3 Tools and Techniques

DG is fundamentally about organisational behaviour, but tools can help with communication, metrics and business glossary development. Requirements must be clearly defined before purchasing a tool.

- **Online Presence / Websites:** For collaboration, communication and sharing documents

- **Business Glossary:** Core DG tool housing agreed-upon definitions and business terms and relates them to data
- **Workflow tools:** Connects processes to documents and is useful in policy admin and issue resolution
- **Document management tools**
- **Data Governance Scorecards:** Collection of metrics to track DG activities. Can be automated

4 Implementation Guidelines

4.1 Organisation and culture

The target of organisational change is **sustainability** (a quality of a process that measures how easy it is for the process to continue to add value).

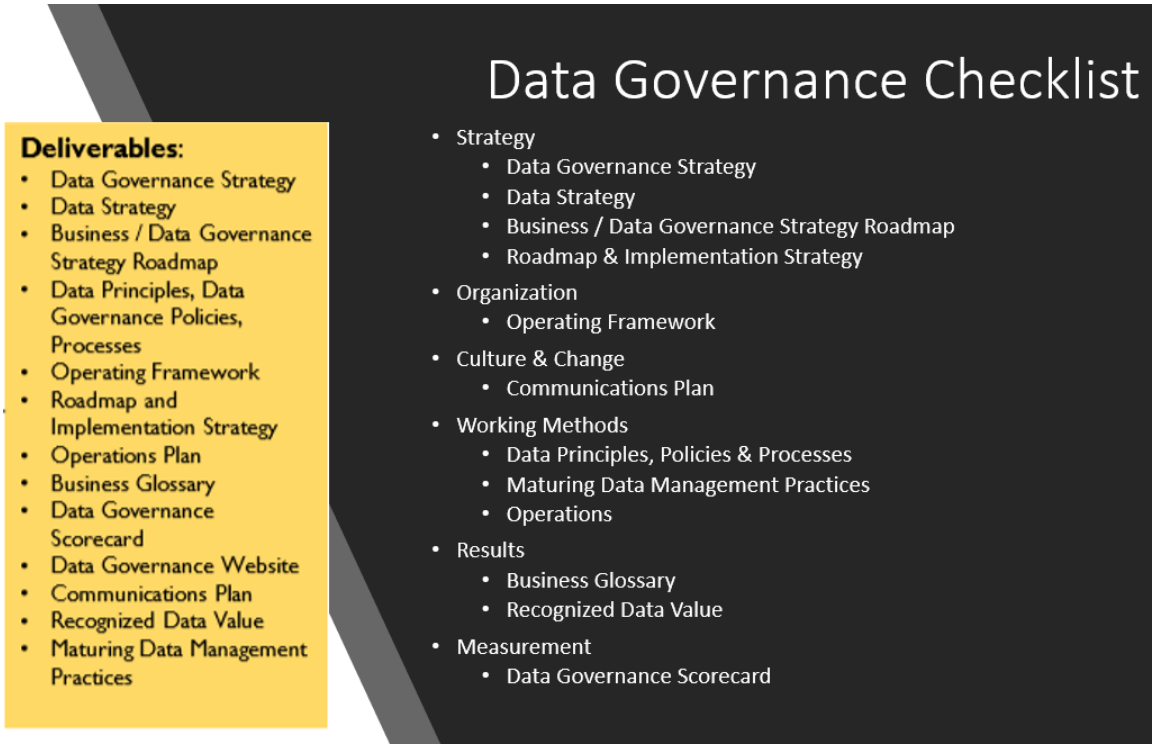
4.2 Adjustment and Communication

- Business / DG strategy map
- DG Roadmap
- Ongoing business case for DG
- DG Metrics

5 Metrics

Data Governance function must be able to measure progress and success and how DG participants have added value to business objectives.

- **Value**: Contributions to business objectives, reduction of risk, improved efficiency
- **Effectiveness**: Achievement of goals and objectives, Stewards using relevant tools, Effectiveness of communication, education and training
- **Sustainability**: Policies and processes working appropriately, Staff conforming to standards and procedures



Knowledge Area Data Governance

Data Governance	Measurement	Policy Framework	Change Management

• Data Sources to be integrated
• Data Quality rules to be enforced
• Conditions of use rules
• Activities to be monitored and the frequency of monitoring
• Priority and response levels of stewardship efforts
• How information is represented to meet stakeholder needs
• Standard approval gates, expectations in RDM and MDM deployment
	• Leading Indicators
	  • Knowledge Area Metric
	• Lagging Indicators
	  • Business SWOTS
	  • Insights
	  • Decision Making
	  • Revenue
	• Data Management SWOTS (Friction)
		• Principles
		• Policies
		• Procedures
		• Standards
		• Guidelines
		• Roles & Responsibilities
		• Policy Relationships
		  • Industry
		  • Organization
		  • Internal
			• Vision
			• Culture
			  • Decision Making
			  • Attitudes
			  • Behaviour
			• Communication
			• Collaboration Structures
			  • Agile

How Data Governance oversees the other Knowledge Areas (Howard Diesel)

Chapter 4
Data Architecture

1 Introduction

The essential components of Data Architecture:

- **Data Architecture Outcomes:** The Data Architecture Artefacts
- **Data Architecture Activities:** To fulfil Data Architecture intentions
- **Data Architecture Behaviour:** Collaboration among roles with an Enterprise view

Data Architecture Is fundamental to Data Management. The vast data of an organisation must be represented at various levels of abstraction so that it can be understood for management to make decisions.

An organisation's Data Architecture consists of:

- Master design documents at different abstraction
- Formal **Enterprise Data Model** – containing data names, metadata definitions conceptual and logical entities and relationships and business rules

Data Architecture **enables consistent data standardisation and integration** across the enterprise.

Data Architecture artefacts constitute metadata and should be stored and managed in an enterprise architecture artefact repository.

ISO/IEV 42010:2007 Definition:

The fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution.

What is a Data Architecture? Modern Data Architectures Explained - YouTube

1.1 Business Drivers

- Strategic preparation of evolution of products, services and data to take advantage of business opportunities in emerging technologies
- Translate business needs into data and system requirements
- Manage complex data delivery throughout the enterprise
- Facilitate alignment between business and IT
- Act as agents for change, transformation and agility
- These business drivers should influence measures of the value of Data Architecture

Data Architecture is **a bridge between business strategy and technology execution**. Data architects create and maintain organisational knowledge about data and the systems through which it moves, which enables it to maintain data as an asset, and to increase the value it gets from data.

1.2 Data Architecture Outcomes and Practices

Primary Data Architecture outcomes:

- Data storage and processing requirements

- Designs of structure and plans that meet current and long term data requirements of the enterprise

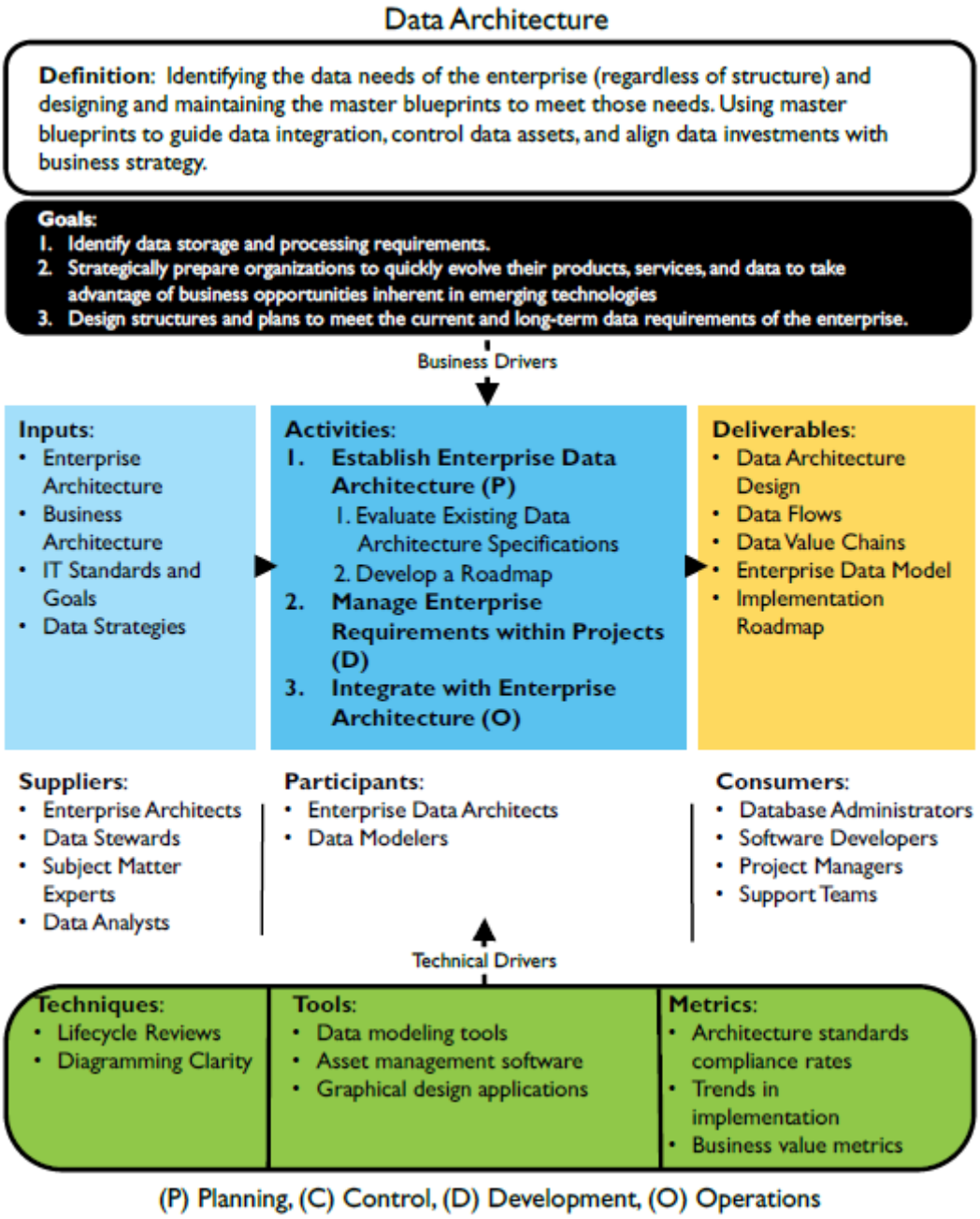

Data Architecture
Definition: Identifying the data needs of the enterprise (regardless of structure) and designing and maintaining the master blueprints to meet those needs. Using master blueprints to guide data integration, control data assets, and align data investments with business strategy.
Goals:
1. Identify data storage and processing requirements.
2. Strategically prepare organizations to quickly evolve their products, services, and data to take advantage of business opportunities inherent in emerging technologies
3. Design structures and plans to meet the current and long-term data requirements of the enterprise.
Business Drivers
Inputs:
• Enterprise Architecture
• Business Architecture
• IT Standards and Goals
• Data Strategies
Activities:
1. Establish Enterprise Data Architecture (P)
1. Evaluate Existing Data Architecture Specifications
2. Develop a Roadmap
2. Manage Enterprise Requirements within Projects (D)
3. Integrate with Enterprise Architecture (O)
Deliverables:
• Data Architecture Design
• Data Flows
• Data Value Chains
• Enterprise Data Model
• Implementation Roadmap
Suppliers:
• Enterprise Architects
• Data Stewards
• Subject Matter Experts
• Data Analysts
Participants:
• Enterprise Data Architects
• Data Modelers
Consumers:
• Database Administrators
• Software Developers
• Project Managers
• Support Teams
Technical Drivers
Techniques:
• Lifecycle Reviews
• Diagramming Clarity
Tools:
• Data modeling tools
• Asset management software
• Graphical design applications
Metrics:
• Architecture standards compliance rates
• Trends in implementation
• Business value metrics
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 21 Context Diagram: Data Architecture

To reach goals Data Architects define and maintain specifications that:

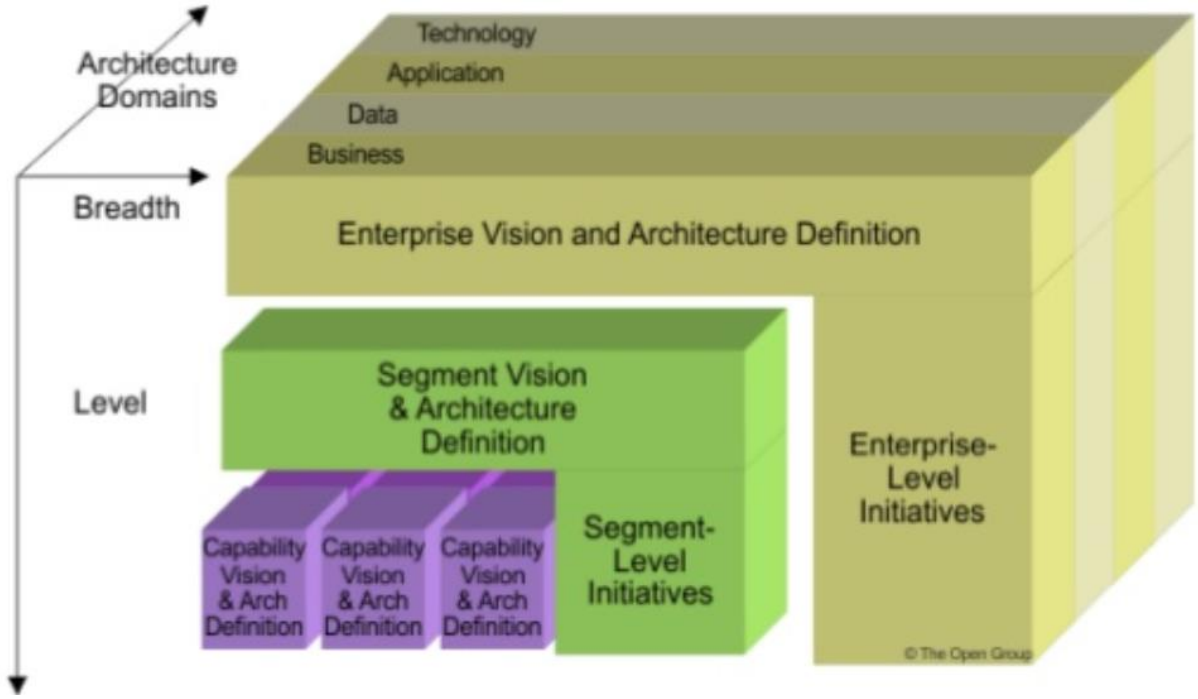
- Define the current state of the organisation
- Provide standard business vocabulary for data and components
- Align Data Architecture with enterprise strategy and business architecture
- Express strategic data requirements
- Outline high level designs to meet these requirements
- Integrate with overall enterprise architecture roadmap

1.3 Essential Concepts

1.3.1 Enterprise Architecture Domains

Domain	Enterprise Business Architecture	Enterprise Data Architecture	Enterprise Applications Architecture	Enterprise Technology Architecture
Purpose	To identify how an enterprise creates value for customers and other stakeholders	To describe how data should be organized and managed	To describe the structure and functionality of applications in an enterprise	To describe the physical technology needed to enable systems to function and deliver value
Elements	Business models, processes, capabilities, services, events, strategies, vocabulary	Data models, data definitions, data mapping specifications, data flows, structured data APIs	Business systems, software packages, databases	Technical platforms, networks, security, integration tools
Dependencies	Establishes requirements for the other domains	Manages data created and required by business architecture	Acts on specified data according to business requirements	Hosts and executes the application architecture
Roles	Business architects and analysts, business data stewards	Data architects and modelers, data stewards	Applications architects	Infrastructure architects

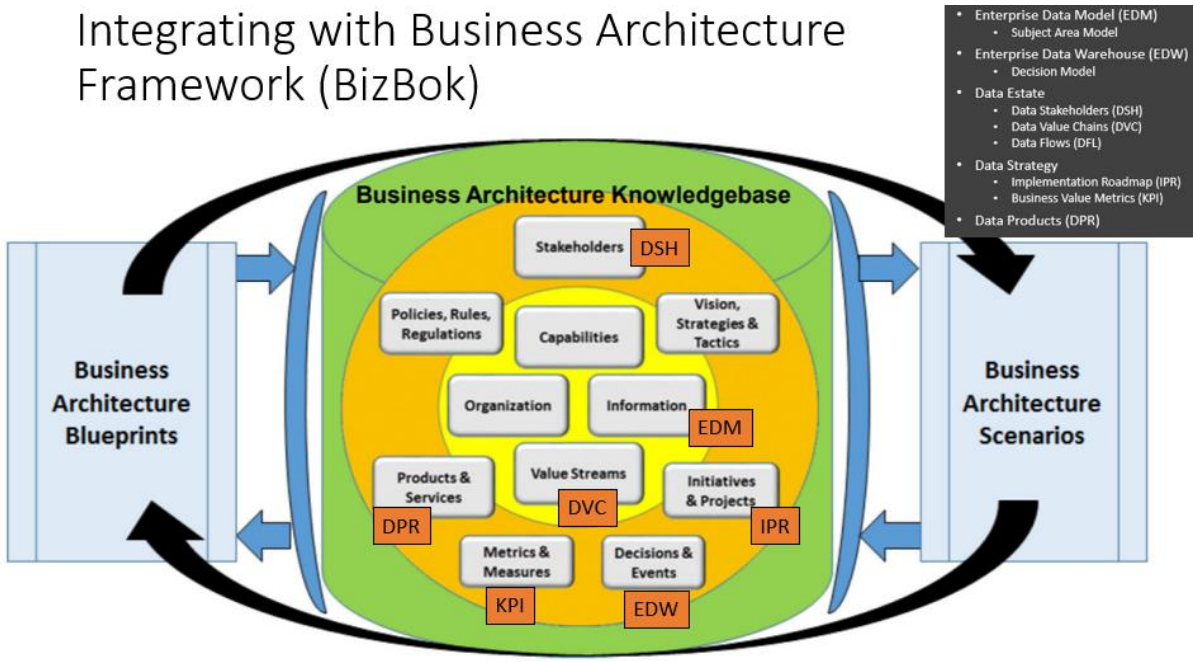
Table 6 Architecture Domains



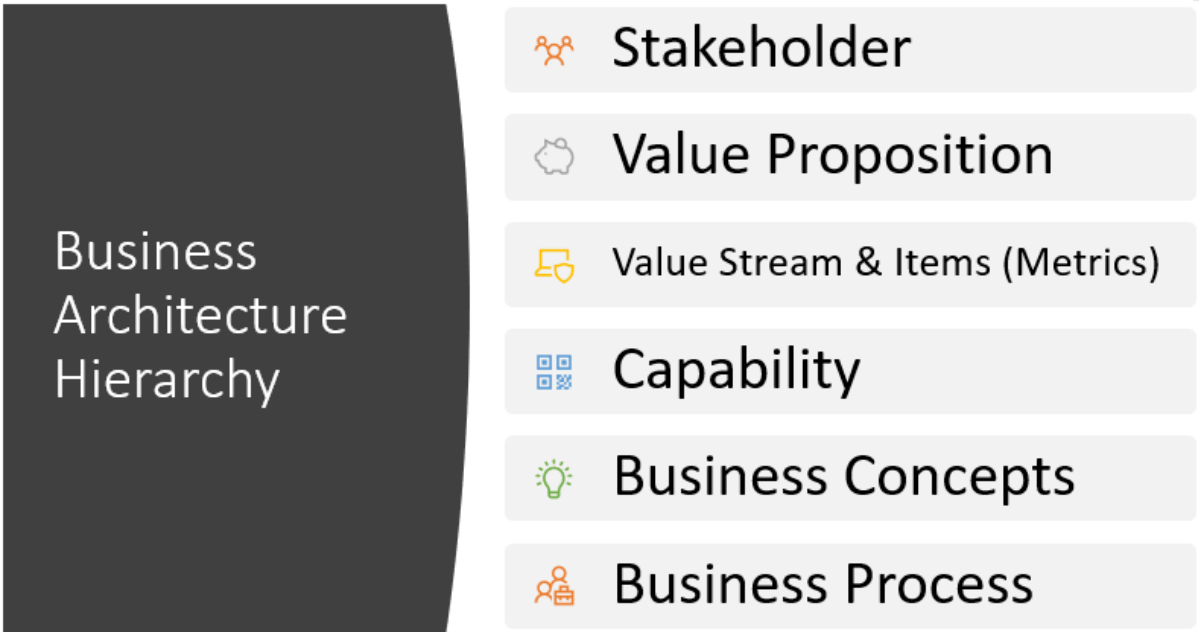
TOGAF – The Open Group Architecture Framework

These next few slides are not in the DMBOK, but are included here to provide context and show the relationships between the business and data architectures.

DMBOK meets BizBok - *This diagram shows the alignment between business and data. Orange blocks represent Data Architecture deliverables in line with the business domains*



Alignment of Data Architecture Deliverables with the BizBoK Framework



The totality of physical data can be described as the **data landscape**, showing where the organization data is located in relation to the core business concepts in the organization.

This example shows the **core business concepts / business landscape** of a central bank, with both the business data value chain delivering value to customers, and the supporting data value chain.

Business Architecture has 2 main activities:

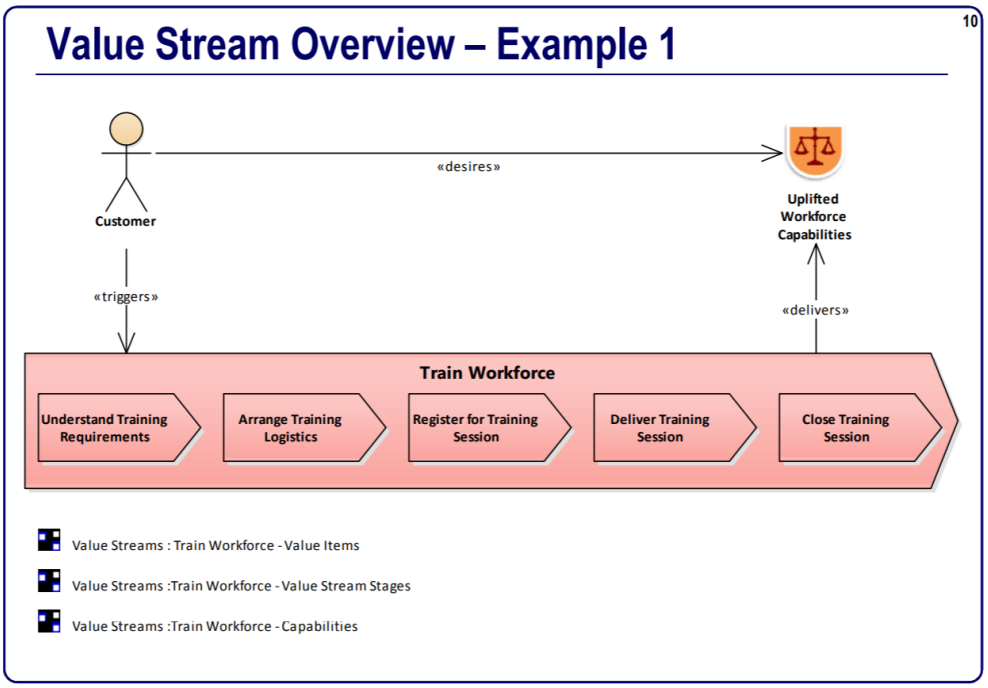
1. **Establish Enterprise Data Architecture**, and
2. **Integrate with Enterprise Architecture** and this integrate step needs to be in line with the business architecture

This helps to answer the questions of:

- What data does the enterprise manage?
- Where is it located?

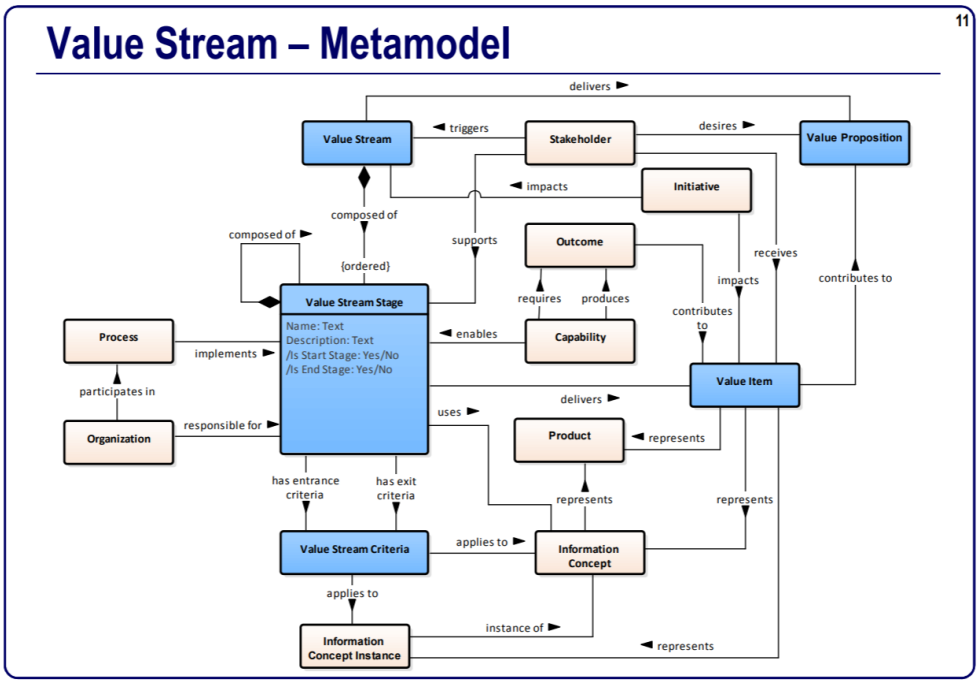
Components / deliverables of the Business Architecture (which need to map into the data architecture)

- **Stakeholders** are the people involved in the business, whether suppliers, customers or the people in the organisation. There are business and data stakeholders
- A **Value Propositions** is an innovation, service, or feature intended to make a company or product attractive to customers, and consists of
  - **Value streams** are *the sequence of activities necessary to deliver a product, service or experience to a customer, internal or external*
    from the initial request through realization of value by the customer –
    eg of order to fulfilment value stream in a fast food restaurant :
    review the menus -> place order -> submit the coupon -> pay the bill -> pick up the food -> enjoy & eat
  - **Data Value Items,** each activity in the value stream
  - **Metrics**
- **Business Capability**: WHAT you need to be able to do to deliver value (the primary business functions of an enterprise and the pieces of the organization that perform those functions)
- Business concepts translate to information concepts = **data required**
- Business process is the more detailed set of steps showing **how** to deliver the value stream

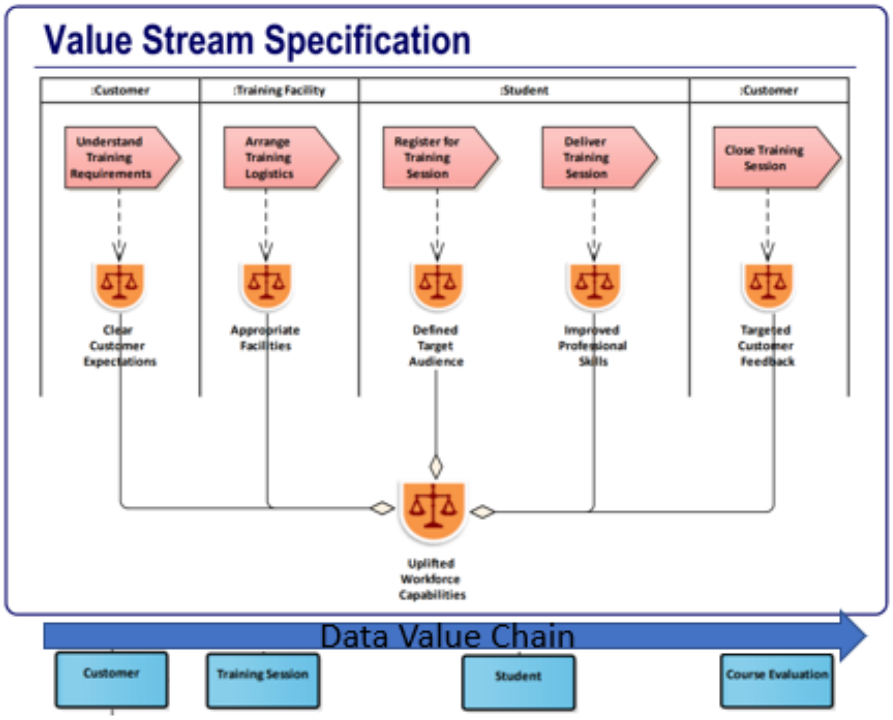


Example of a Training business: The stakeholder requirement is to have an uplifted workforce.

- **Business Value stream:** (pink process flow)
- **Train Workforce**, is the deliverable from the Business Architecture
- *The value stream shows **the sequence of activities necessary*** to deliver value to stakeholder (ie HOW)
- Value streams may be externally or internally triggered, here it is triggered by a customer requesting training

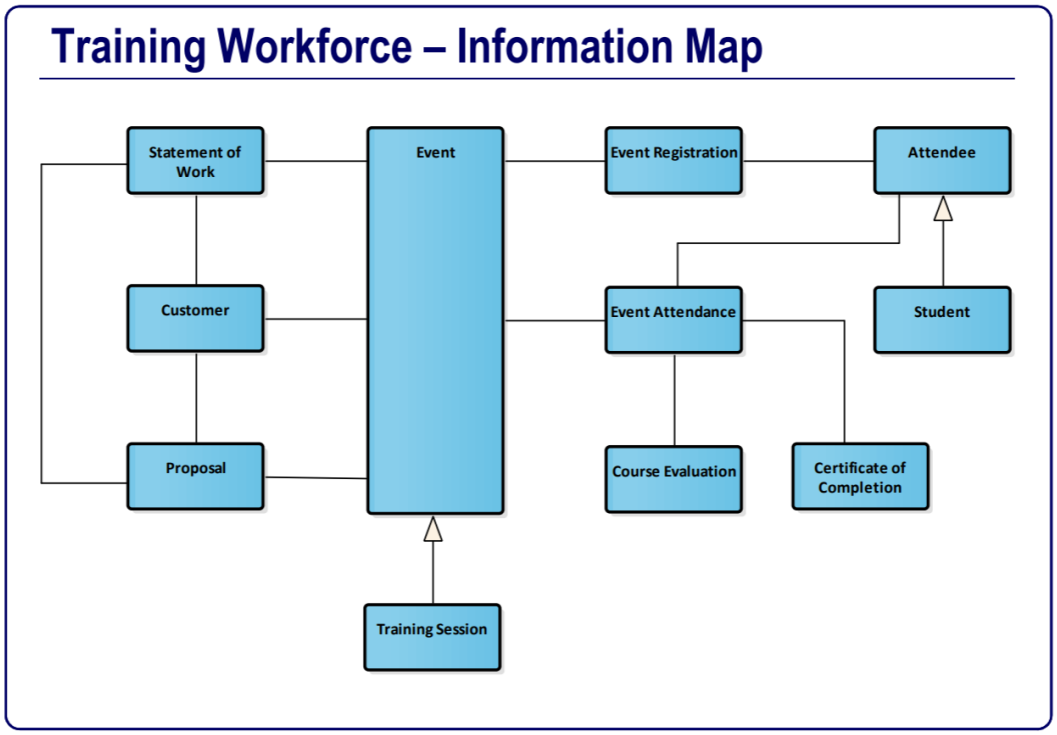


Metamodel stores info about the value stream as a business model

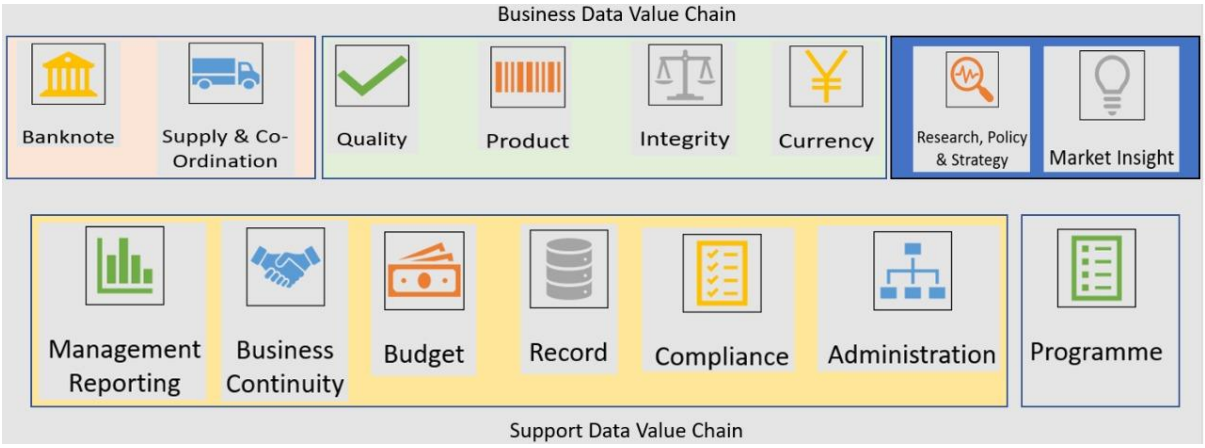


The Business Value Stream is translated into the Value Stream Specification *(which is mapped to the Data Value Chain)*

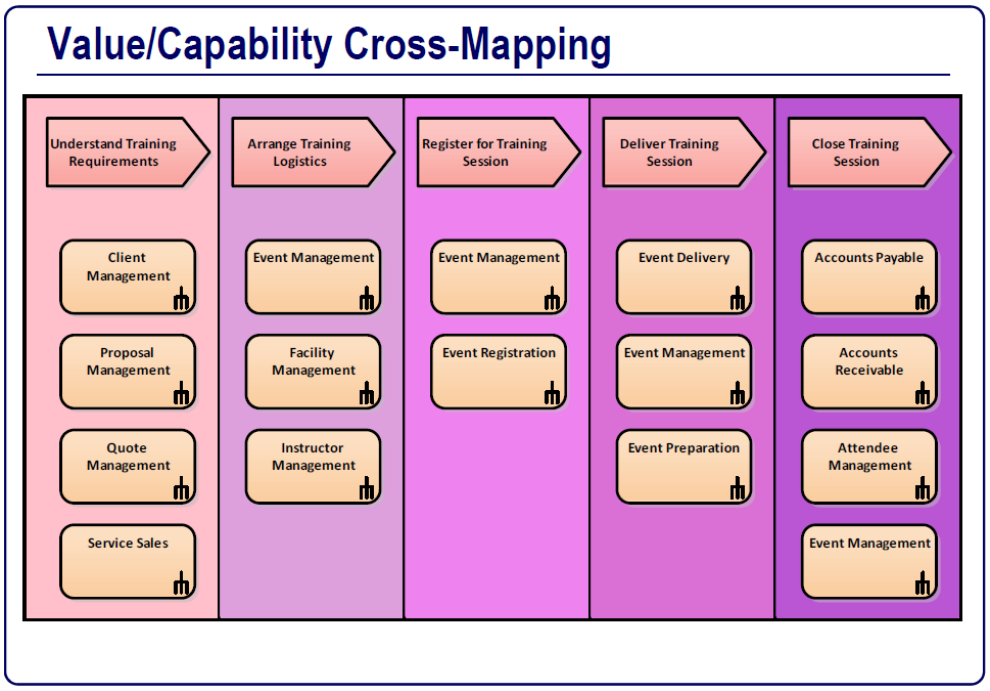
- **Value Stream Specification:**
  - **Business Entities** are the top line / lanes. They are nouns, eg **Customer**, **Training Facility** (the description of the Business Entities at the bottom in the Data Value Chain are constrained by the business model of the training provider e.g., Customer Training Facility is mapped to Modelware's Training Session)
  - **Value Stream items**: (Understand **Training Requirements**, **Arrange Training Logistics** etc). Value stream items are grouped by Business Entity (eg **Register for Training Session** & **Deliver Training Session** are grouped under the **Student** Business Entity)
  - There is a metric (orange scale) used to measure Value stream item. eg **Clear Customer Expectations** measures the Value Stream Item **Understand Training** requirements
  - The metrics above contribute to the **Uplifted Workforce Capabilities** metric provided to the customer
- **Data Value Chain**
  - The improvement of the value of data aligned to the business value chain. The data value chain must support the ongoing measurement and improvement of the business value chain
  - Data Value Chain (consists of Business entities, eg Customer, Training Section)
  - **Course evaluation** is the most important to measure success – do we have an uplifted workforce? Outcomes
  - DVC is critical data which builds. Each element enriches the next element. Each segment produces another value item. All comes together as total value
- Business value streams align to data value chain (helps to identify major areas in EDM)



The Enterprise Subject Area Model



The Data Landscape of a Central Bank Showing Major Subject Areas



Business Capability Map is another deliverable of the Business Architecture

The capabilities will give us an idea of the entities we will be mapping, for example, we have **Client Management Capability**, therefore we will need to be including a **Client** entity in our Data Model.

1.3.2 Enterprise Architecture Frameworks

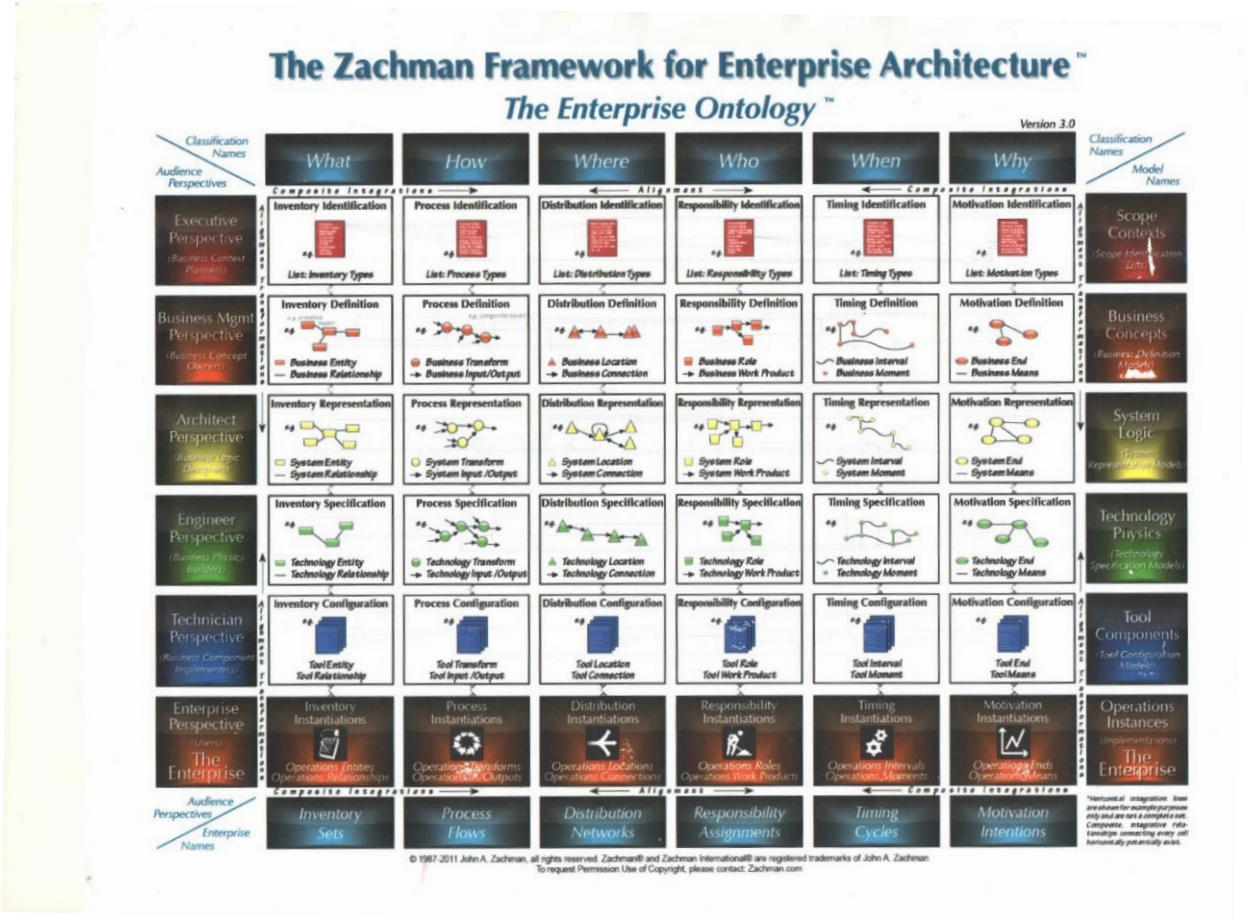
Provide a framework for thinking about and understanding architecture.

1.3.2.1 Zachman Framework for Enterprise Architecture

Developed by John A. Zachman in the 1980s. An ontology. The 6 x 6 matrix shows the models required to describe an enterprise, and the relationships between them. Each cell represents a unique type of design artefact defined by the intersection of row and column.

	What	How	Where	Who	When	Why	
Executive	Inventory Identification	Process Identification	Distribution Identification	Responsibility Identification	Timing Identification	Motivation Identification	Scope Context
Business Management	Inventory definition	Process Definition	Distribution Definition	Responsibility Definition	Timing Definition	Motivation Definition	Business Concepts
Architect	Inventory Representation	Process Representation	Distribution Representation	Responsibility Representation	Timing Representation	Motivation Representation	System Logic
Engineer	Inventory Specification	Process Specification	Distribution Specification	Responsibility Specification	Timing Specification	Motivation Specification	Technology Physics
Technician	Inventory Configuration	Process Configuration	Distribution Configuration	Responsibility Configuration	Timing Configuration	Motivation Configuration	Tool Components
Enterprise	Inventory Instantiations	Process Instantiations	Distribution Instantiations	Responsibility Instantiations	Timing Instantiations	Motivation Instantiations	Operational Instances
	Inventory Sets	Process Flows	Distribution Networks	Responsibility Assignments	Timing Cycles	Motivation Intentions	

Figure 22 Simplified Zachman Framework



Zachman Framework (courtesy of John Zachman)

Columns: Communication interrogatives

- **What** (the inventory column): Entities used to build the architecture
- **How** (the process column: Activities performed

- **Where** (the distribution column): Business location and technology location
- **Who** (the responsibility column): Roles and organisations
- **When** (the timing column): Intervals, events, cycles and schedules
- **Why** (the motivation column): Goals, strategies and means

Rows: Reification transformations – the steps necessary to translate an abstract idea into a concrete instance from different perspectives (planner, owner, designer, builder, implementer and user). Each perspective has a different relation to the What column:

- The **executive** perspective – Business context: Lists of business elements defining scope
- The **business management** perspective – Business concepts: Relationships between business concepts
- The **architect** perspective – Business logic: System logical models
- The **engineer** perspective – Business physics: Physical models optimised for technology, people, costs and timeframes
- The **technician** perspective – Component assemblies: Technology specific
- The **user** perspective – Operations classes: Actual functioning instances used by workers

1.3.3 Enterprise Data Architecture

Enterprise Data Architecture descriptions include:

- **Enterprise Data Model (EDM**): A holistic, enterprise-level, implementation-independent conceptual or logical data model providing a common consistent view of data across the enterprise
- **Data flow design:** Defines requirements and master blueprint for storage and processing across databases, applications, platforms and networks (the components)

Both need to be reflected in current state (architecture perspective) and transition state (project perspective).

1.3.3.1 Enterprise Data Model

Can be a stand-alone artefact or may be composed of data models from different perspectives or levels of detail. An EDM includes both universal (Enterprise-wide Conceptual and Logical Models) and application or project specific data models, along with definitions, specifications, mappings and business rules.

Industry standard is a start but time and effort must also be invested to build and maintain the EDM.

- Conceptual overview over the enterprise's subject areas
- Views of entities and relationships for each subject area
- Detailed, partially attributed logical views for the same subject area
- Logical and physical models specific to an application or project

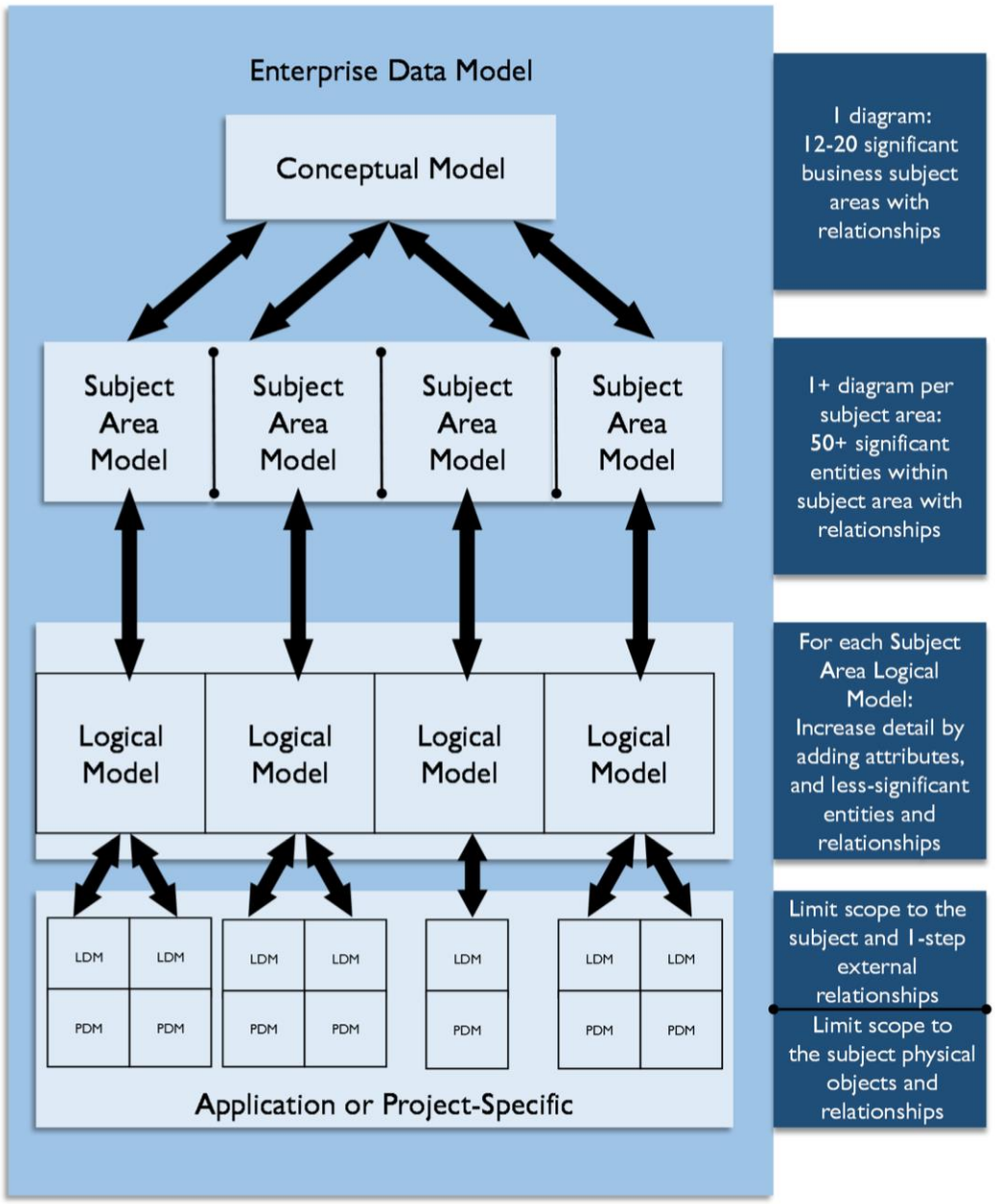

Enterprise Data Model
Conceptual Model
Subject Area Model
Subject Area Model
Subject Area Model
Subject Area Model
Logical Model
Logical Model
Logical Model
Logical Model
LDM
PDM
Application or Project-Specific
1 diagram: 12-20 significant business subject areas with relationships
1+ diagram per subject area: 50+ significant entities within subject area with relationships
For each Subject Area Logical Model: Increase detail by adding attributes, and less-significant entities and relationships
Limit scope to the subject and 1-step external relationships
Limit scope to the subject physical objects and relationships

Figure 23 Enterprise Data Model

Different types of models are related vertically and horizontally.

The Enterprise Data Model is built up by the combination of Subject Area Models:

- **Top Down approach**: Form Subject Areas then populate them with models
- **Bottom up Approach:** Subject Area structure is based on existing data models
- **Middle out:** Combination of both approaches starting from a middle layer

The Subject Area Discriminator (the way subject areas are formed) must be consistent throughout the enterprise data model:

- **Funding:** Systems portfolios
- **Organisational**: Data governance structure and data ownership
- **Business value chains**: Top-level processes
- **Using business capabilities:** See training organisation example above

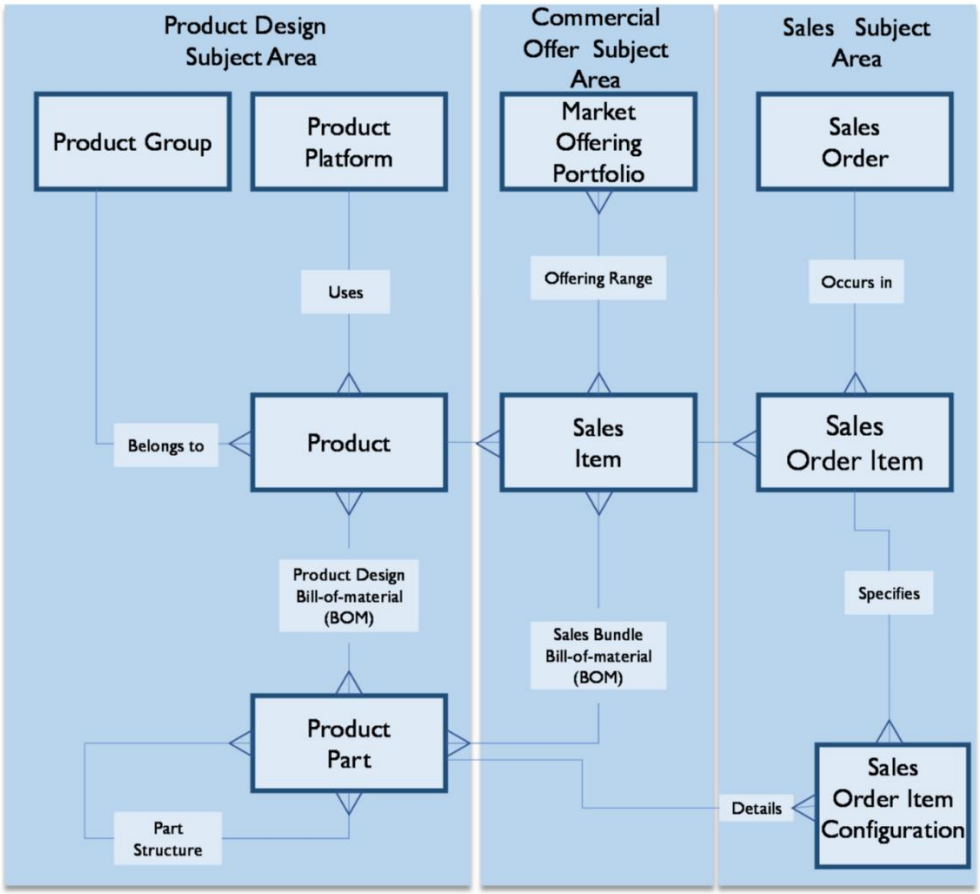


Figure 24 Subject Area Models Diagram Example

1.3.3.2 Data Flow Design

Data flows are data lineage documentation that depicts how data moves end-to-end through business processes and systems, where it originated, is stored and used and how it transforms.

Data flows can be documented at different levels of detail: Subject Area, Business entity or attribute level.

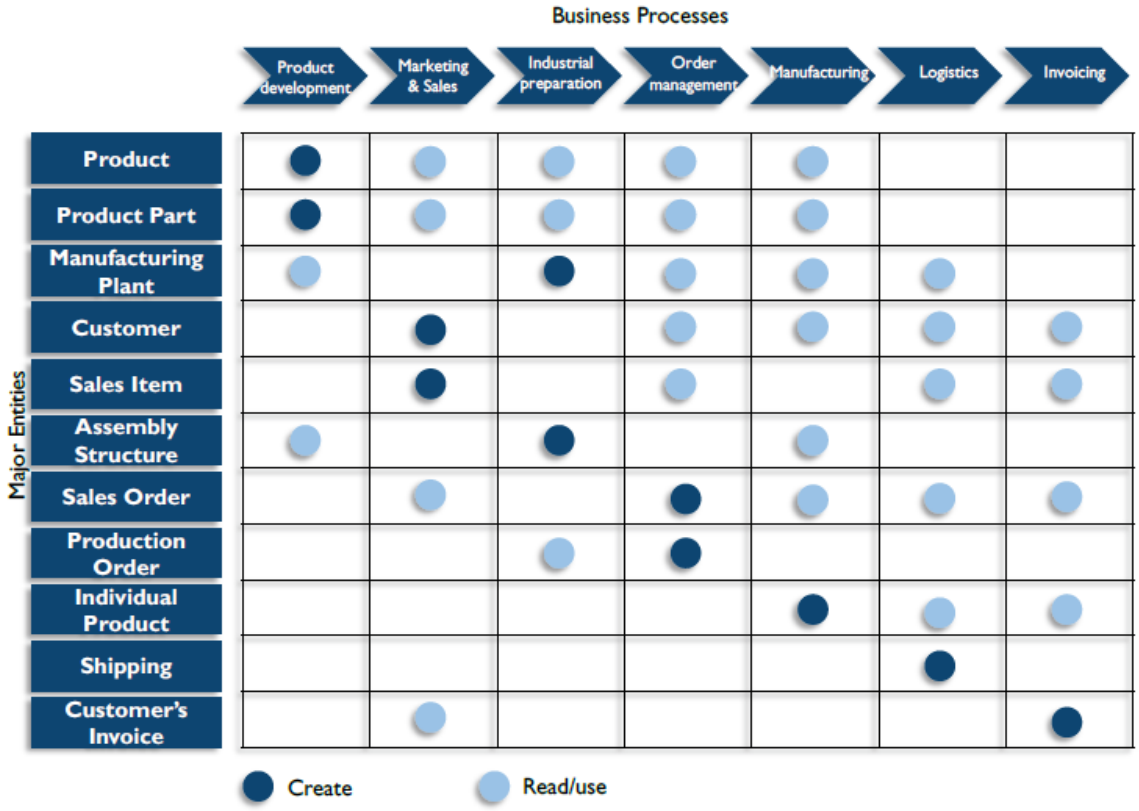


Figure 25 Data Flow Depicted in a Matrix

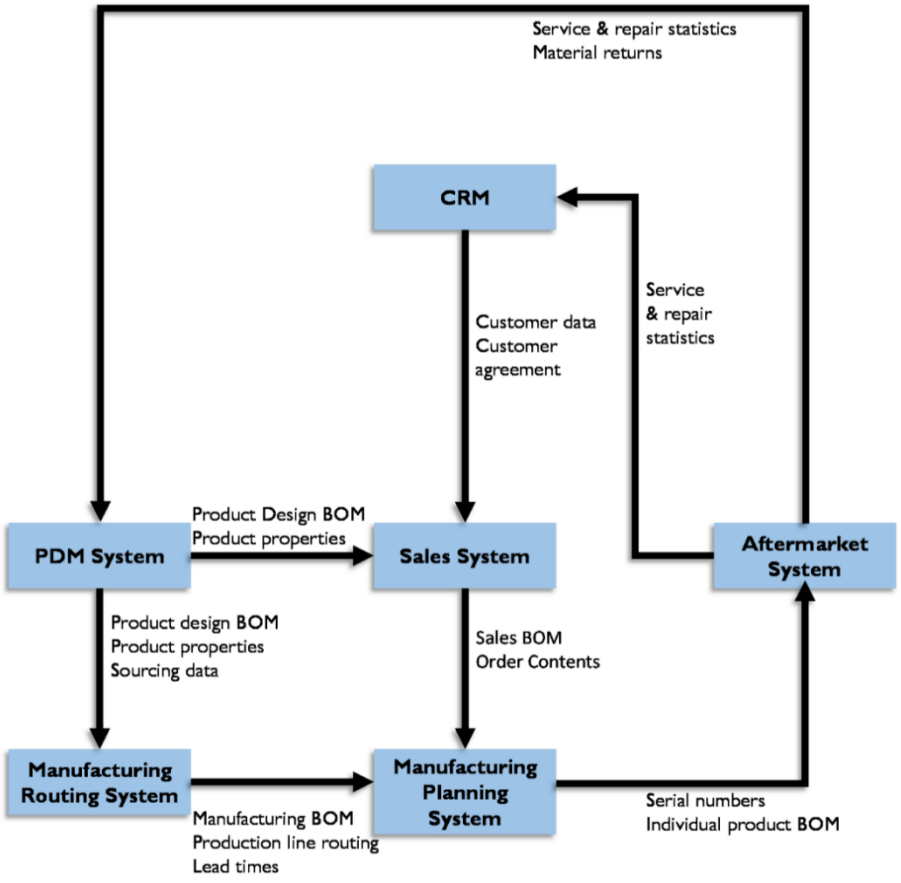


Figure 26 Data Flow Diagram Example

2 Activities

Two approaches:

- **Quality-oriented**: Focus on execution within business and IT structures. Unless architecture is managed it will deteriorate. Architectural improvements are incremental
- **Innovation-oriented:** Focus on transforming business and IT to address new expectations and opportunities. Requires interaction with business development representatives and designers

2.1 Establish Data Architecture Practice

A Data Architecture practice includes the following work streams:

- **Strategy:** Select frameworks, state approaches, develop roadmap
- **Acceptance and culture:** Inform and motivate changes in behaviour
- **Organisation:** Assign Data Architecture accountabilities and responsibilities
- **Working methods:** Define best practices and perform Data Architecture work within development projects, in coordination with Enterprise Architecture
- **Results:** Produce Data Architecture artefacts within an overall roadmap

Enterprise Architecture influences scope boundaries of projects/system releases:

- Defining project data requirements
- Reviewing project data designs
- Determining data lineage impact
- Data replication control
- Enforcing Data Architecture standards
- Guide data technology and renewal decisions

2.1.1 Evaluate Existing Data Architecture Specifications

Evaluate existing documentation for accuracy, completeness and level of detail. Update if necessary.

2.1.2 Develop a Roadmap

Describes the architecture's 3 to 5-year development path, with business requirements, consideration of actual conditions and technical assessments. Must be integrated into the enterprise architecture roadmap. Include milestones, resources needed, cost estimations, and divided into work streams.

The Enterprise Data Architecture can be formed by resolving input and output data flows in the chain of dependencies between business capabilities.

Start with the most independent business capabilities and end with those most dependent on other activities. The following diagram has the lowest dependency on the top.

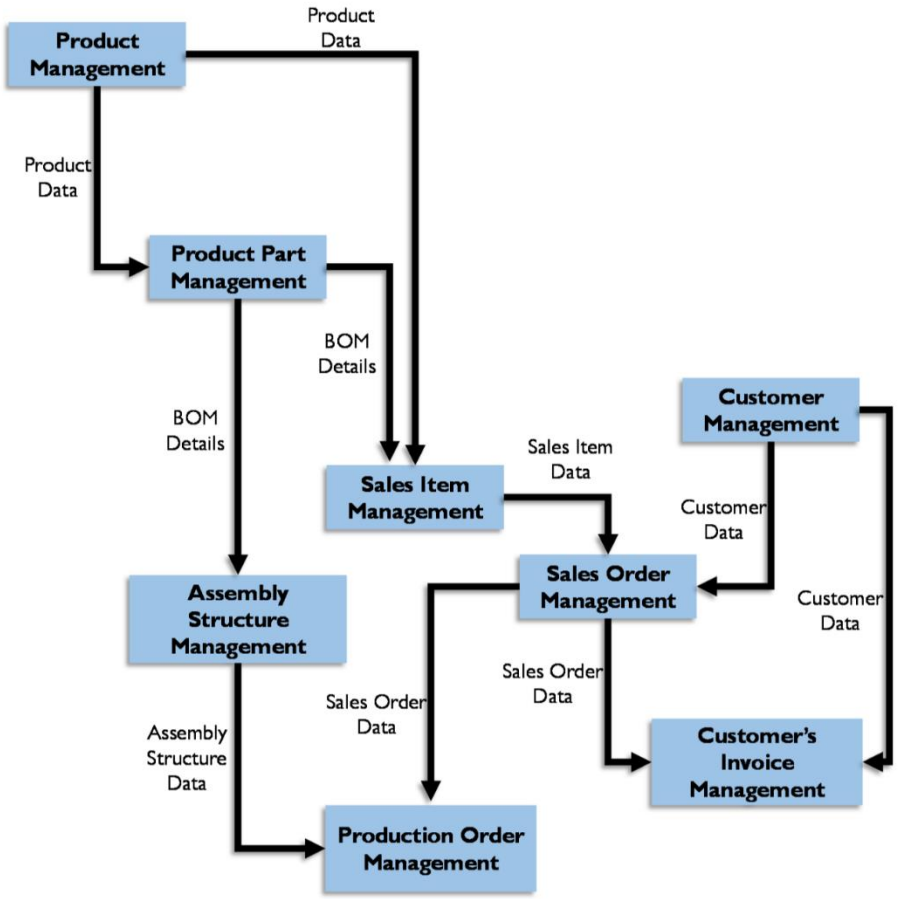


Figure 27 The Data Dependencies of Business Capabilities

2.1.3 Manage Enterprise Requirements within Projects

Enterprise Data Architecture project-related activities include:

- **Define scope:** Ensure the scope and interface are aligned with Enterprise Data Model. Identify components that can be reused, and down-stream dependencies
- **Understand business requirements**
- **Design:** Form detailed target specifications. Look for shareable constructs in the enterprise logical data model. Review and use technology standards
- **Implement:**
    - **When buying:** Reverse engineer purchased applications and map against data structure
    - **When reusing data:** Map application data models against common data structures and new and existing processes to understand CRUD operations. Enforce use of authoritative data
    - **When building:** Data storage according to data structure

The role of Enterprise Data Architects and the process of building architectural activities into projects depends on the development methodologies:

- **Waterfall methods:** Construct systems in sequential phases as part of an overall design
- **Incremental methods:** Learn and construct in gradual steps. creates prototypes based on vague overall requirements
- **Agile, iterative methods:** Learn, construct and test in discrete delivery packages (Sprints)

2.2 Integrate with Enterprise Architecture

Integrate Enterprise Data Architecture matters with project portfolio management as funded projects drive architecture priorities and Data Architecture can influence the scope of projects.

3 Tools

- **Data Modelling Tools:** Include lineage and relation tracking to manage linkages between models
- **Asset Management Software:** Used to inventory systems, describe their content and track the relationships between them
- **Graphical Design Applications:** To create architectural design diagrams and other architectural artefacts

4 Techniques

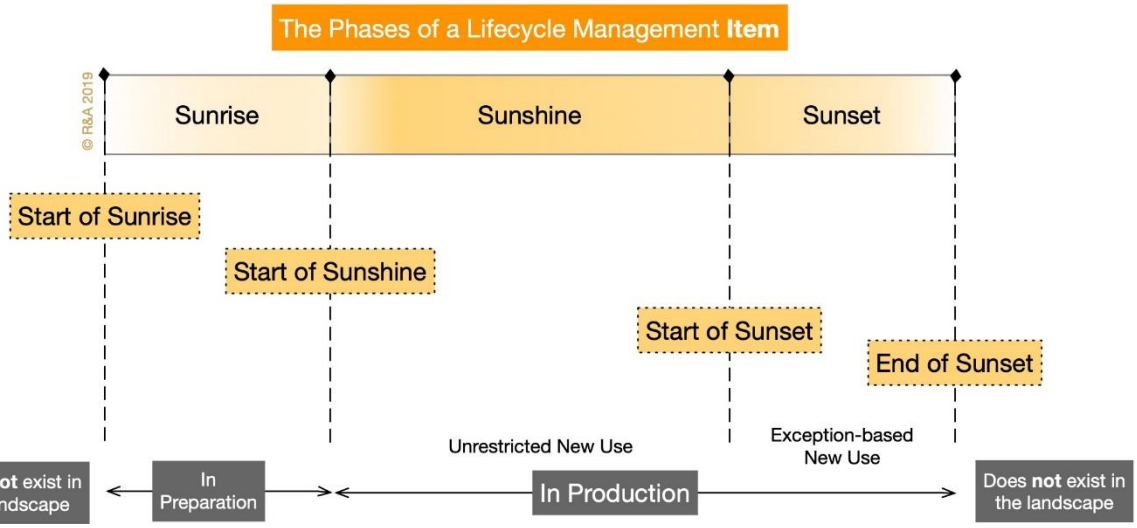
4.1 Lifecycle Projections

Architecture designs can be:

- Aspirational and future-looking
- Implemented and active
- Plans for retirement

What architectural plans represent should be clearly documented:

- **Current:** Products supported and used
- **Deployment period:** Products deployed for use in 1-2 years
- **Strategic period:** Products available in the next 2+ years
- **Retirement:** Retired products, or retirement within 1 year
- **Preferred:** Products preferred for use by most applications
- **Containment:** Limited for use by certain applications
- **Emerging:** Researched and piloted for possible future deployment
- **Reviewed:** Evaluated products and their evaluation results
-



The Phases of a Lifecycle Management Item (Author Unknown)

4.2 Diagramming clarity

Models and diagrams must conform to an established set of visual conventions:

- **Clear and consistent legend:** Identify all objects and lines and placed in the same spot in all diagrams.
- **Match between all diagram objects and the legend:** Not all legend objects need appear on diagram
- **Clear and consistent line direction:** Usually left to right. Backward lines must be clear
- **Consistent object attributes:** Differences in size, line thickness and colour should signify something
- **Linear symmetry:** Line up at least half of the objects to improve readability

5 Implementation Guidelines

As Data Architecture is about artefacts, activities and behaviour, Enterprise Data Architecture is about:

- Organising Enterprise Data Architecture teams and forums
- Producing initial versions of Data Architecture artefacts such as enterprise data model, enterprise wide data flow and road maps
- Forming and establishing a data architectural way of working in development projects
- Creating organisation wide awareness of the value of Data Architecture efforts

A Data Architecture implementation should include at least 2 of the above.

Data models and other Data Architecture artefacts are captured within development projects and are then standardised and maintained by architects. There will be more architectural work in early projects which may need special architectural funding.

Enterprise Data Architecture evolves incrementally in a solution-oriented culture using agile development.

Enterprise Data Architecture starts with Master Data areas in need of improvement in planned development projects, and expands to include business and other data.

5.1 Readiness / Risk Assessment

More risks than other projects, especially during an organisation's first attempt:

- Lack of management support
- No proven record of accomplishment
- Apprehensive sponsor
- Counter-productive executive decisions
- Culture shock
- Inexperienced project leader
- Dominance of a one-dimensional view

5.2 Organisation and Cultural change

The ability of an organisation to adopt Data Architecture practices depends on several factors:

- Cultural receptivity to architectural approach
- Organisation recognises data as a business asset, not just an IT concern

- Ability to let go of a local perspective and adopt an enterprise perspective on data
- Ability to integrate architectural deliverables into project methodology
- Level of acceptance of formal data governance
- Ability to look holistically at the enterprise

6 Data Architecture Governance

Enterprise Data Architecture and the Data Governance organisation must be well aligned. A data steward and a data architect should be assigned to each subject area, even to each entity within, as Data Architecture activities support the alignment and control of data. Business event subject areas should be aligned with business processes governance as each event entity usually corresponds to a business process.

Data Architecture governance activities include:

- **Overseeing projects:** Projects comply with required Data Architecture activities, use architectural assets and are implemented according to data architectural standards
- **Managing architectural designs, lifecycle and tools:** Designs must be defined, evaluated and maintained
- **Defining standards**
- **Creating data-related artefacts**

6.1 Metrics

Data Architecture metrics may be monitored annually for business customer satisfaction:

- **Architecture standard compliance rate:** How far projects comply with established data architectures
- **Implementation trends:** The degree to which enterprise architecture has improved the organisation's ability to implement projects along at least two lines:
    - **Use/reuse/replace/retire measurements:** Proportion of new architectural artefacts to reused, replaced or retired artefacts
    - **Project execution efficiency measurements:** Measure lead times for projects and their resource costs for delivery improvements with reusable artefacts and guiding artefacts
- **Business value measurements:** Track progress towards expected business benefits
    - **Business agility improvements:** Account for the benefits of lifecycle improvements or the cost of delay

- **Business quality:** Measure whether business cases are fulfilled and projects deliver changes leading to business improvements
- **Business operation quality:** Measure of improved efficiency and accuracy
- **Business environment improvements**

Chapter 5
Data Modelling and Design

1 Introduction

Data Modelling is the process of discovering, analysing and scoping data requirements, and then representing and communicating these data requirements in a precise form called the data model.

Six most commonly used schemes to represent data are:

- Relational
- Dimensional
- Object-Oriented
- Fact-Based
- Time-Based
- NoSQL

Models exist at three levels of detail:

- Conceptual
- Logical
- Physical

Each model contains a set of components such as entities, relationships attributes, facts and keys. Data Models contain Metadata uncovered during the modelling process which is essential to other data management functions.

1.1 Business Drivers

Data models:

- Provide common vocabulary around data
- Capture and document explicit knowledge about the organisation's data and systems
- Primary communications tool during projects
- Provide the starting point for customisation, integration or replacement of an application

1.2 Goals and Principles

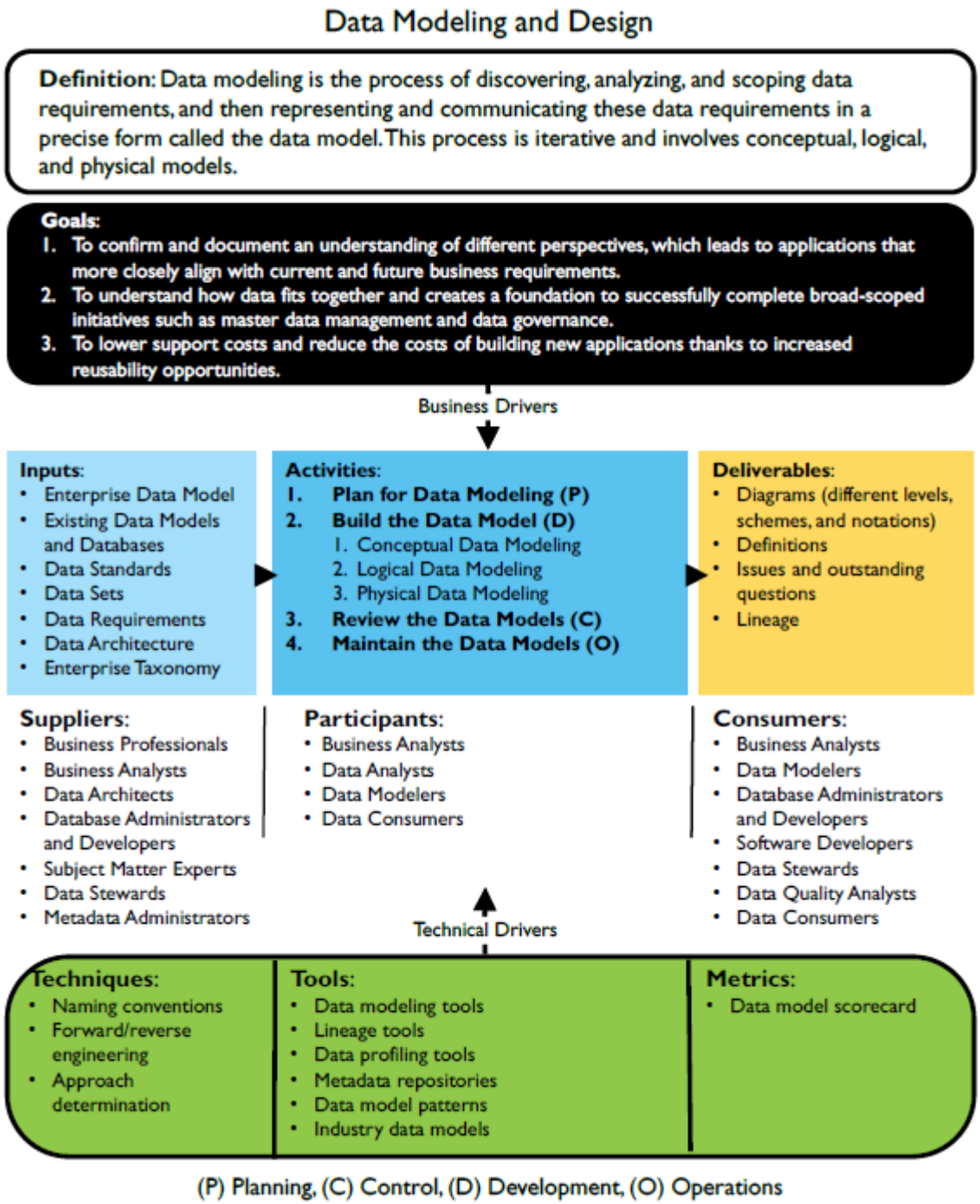
Goals:

1. To confirm and document an understanding of different perspectives, which leads to applications that more closely align with current and future business requirements.
2. To understand how data fits together and creates a foundation to successfully complete broad-scoped initiatives such as master data management and data governance.
3. To lower support costs and reduce the costs of building new applications thanks to increased reusability opportunities.

Confirming and documenting understanding from different perspectives:

- **Formalisation**: A concise definition of data structures and relationships, how data is affected by implemented business rules. A disciplined structure reduces the occurrence of data anomalies

- **Scope definition**: Explain boundaries for data context in packages, projects or existing systems
- **Knowledge retention/documentation**: Preserves corporate memory regarding a system by capturing knowledge in an explicit form. Documentation for future projects. Understand the implications of modifications. Data models are reusable maps


Data Modeling and Design
Definition: Data modeling is the process of discovering, analyzing, and scoping data requirements, and then representing and communicating these data requirements in a precise form called the data model. This process is iterative and involves conceptual, logical, and physical models.
Goals:
1. To confirm and document an understanding of different perspectives, which leads to applications that more closely align with current and future business requirements.
2. To understand how data fits together and creates a foundation to successfully complete broad-scoped initiatives such as master data management and data governance.
3. To lower support costs and reduce the costs of building new applications thanks to increased reusability opportunities.
Business Drivers
Inputs:
• Enterprise Data Model
• Existing Data Models and Databases
• Data Standards
• Data Sets
• Data Requirements
• Data Architecture
• Enterprise Taxonomy
Activities:
1. Plan for Data Modeling (P)
2. Build the Data Model (D)
1. Conceptual Data Modeling
2. Logical Data Modeling
3. Physical Data Modeling
3. Review the Data Models (C)
4. Maintain the Data Models (O)
Deliverables:
• Diagrams (different levels, schemes, and notations)
• Definitions
• Issues and outstanding questions
• Lineage
Suppliers:
• Business Professionals
• Business Analysts
• Data Architects
• Database Administrators and Developers
• Subject Matter Experts
• Data Stewards
• Metadata Administrators
Participants:
• Business Analysts
• Data Analysts
• Data Modelers
• Data Consumers
Consumers:
• Business Analysts
• Data Modelers
• Database Administrators and Developers
• Software Developers
• Data Stewards
• Data Quality Analysts
• Data Consumers
Technical Drivers
Techniques:
• Naming conventions
• Forward/reverse engineering
• Approach determination
Tools:
• Data modeling tools
• Lineage tools
• Data profiling tools
• Metadata repositories
• Data model patterns
• Industry data models
Metrics:
• Data model scorecard
(P) Planning, (C) Control, (D) Development, (O) Operations

- Figure 28 Context Diagram: Data Modelling and Design

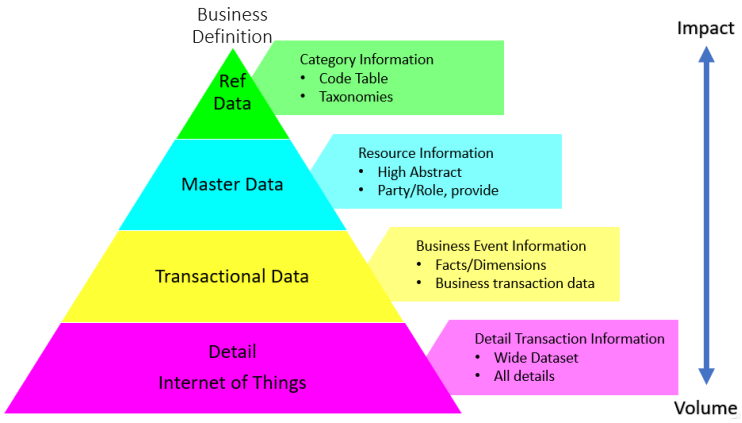
1.3 Essential Concepts

1.3.1 Data Modelling and Data Models

A model consists of diagrams of standard symbols which represent something that exists or something to be made. A data model describes the organisation's data as it is, or as how it wants it to be. Data models are the main medium to communicate data requirements from business to IT and within IT.

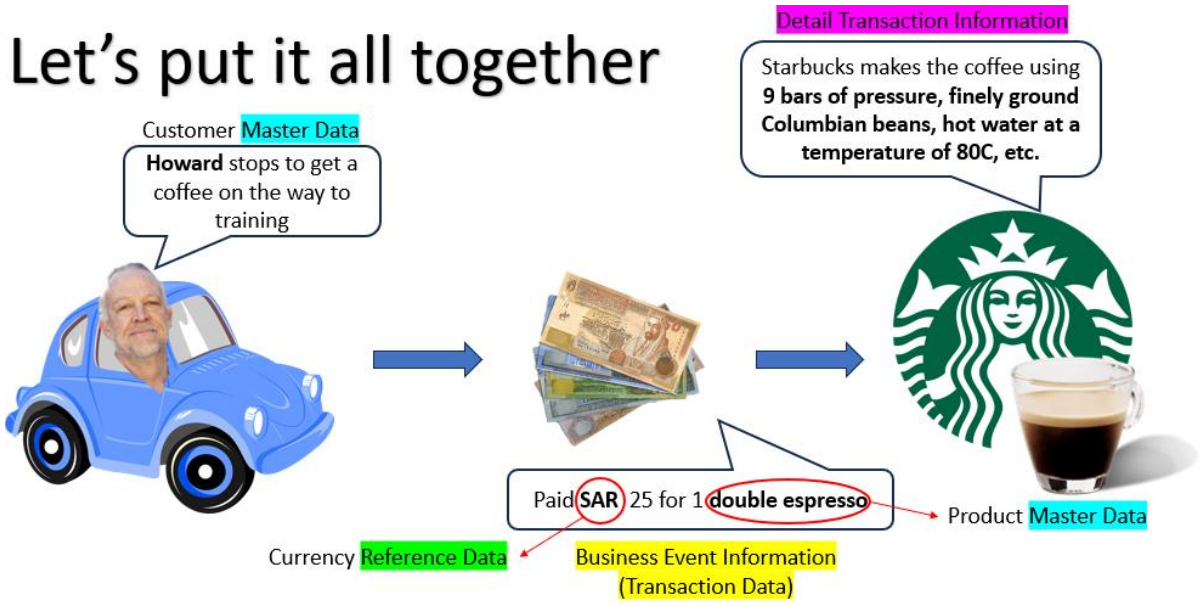
1.3.2 Types of Data That are Modelled

- **Category Information**: Data used to classify and define things (customers classified by market segment or products classified by colour)
- **Resource Information**: Master and Reference Data: objects needed to conduct operational processes such as Product, Customer, Supplier, Facility, Organization, and Account, Countries, Currencies
- **Business Event Information**: Transaction Data created while operational processes are in progress. Examples include Customer Orders, Supplier Invoices, Cash Withdrawal, and Business Meetings
- **Detail Transaction Information**: POS, social media, Clickstream, IoT event. Large volume and rapidly changing. Usually referred to as Big Data. Internet of Things – sensors etc.



Impact vs Volume:

Not so much reference data, but if its wrong (e.g. currency), it has a big impact on the rest. Huge volumes of detailed transaction / big data
Note that if Business definitions are incorrect (Metadata), that has a huge impact on the rest.


Let's put it all together
Customer Master Data
Howard stops to get a coffee on the way to training
Detail Transaction Information
Starbucks makes the coffee using 9 bars of pressure, finely ground Columbian beans, hot water at a temperature of 80C, etc.
Paid SAR 25 for 1 double espresso
Product Master Data
Currency Reference Data
Business Event Information (Transaction Data)

In Riyadh, Howard stops at a Starbucks for a coffee on the way to the training session. He orders a double espresso, and pays SAR 25. He pays using his South African credit card. He is on the loyalty program so they add points.

- **Reference Data**: The currency conversion table from SAR to ZAR. If it was incorrect (e.g. converted to USD) it has a big impact
- **Master Data**: Howard's information is Customer Master Data and the double espresso is Product Master Data. There is more Master Data than Reference Data
- **Business Event Information**: The data generated by this sales transaction. Business Intelligence uses this information
- **Detail Transaction Information**: Big data, Internet of Things, etc. In this example it is the sensor information from the barista machine, such as temperature, pressure and grind

1.3.3 Data Model Components

Basic building blocks of Data models:

- Entities
- Relationships
- Attributes
- Domains

1.3.3.1 Entity

An entity is a thing about which and organisation collects information. A noun of the organisation. Questions to ask to identify entities:

Category	Definition	Examples
Who	Person or organization of interest. That is, *Who* is important to the business? Often a 'who' is associated with a party generalization, or role such as Customer or Vendor. Persons or organizations can have multiple roles or be included in multiple parties.	Employee, Patient, Player, Suspect, Customer, Vendor, Student, Passenger, Competitor, Author
What	Product or service of interest to the enterprise. It often refers to what the organization makes or what service it provides. That is, *What* is important to the business? Attributes for categories, types, etc. are very important here.	Product, Service, Raw Material, Finished Good, Course, Song, Photograph, Book
When	Calendar or time interval of interest to the enterprise. That is, *When* is the business in operation?	Time, Date, Month, Quarter, Year, Calendar, Semester, Fiscal Period, Minute, Departure Time
Where	Location of interest to the enterprise. Location can refer to actual places as well as electronic places. That is, *Where* is business conducted?	Mailing Address, Distribution Point, Website URL, IP Address
Why	Event or transaction of interest to the enterprise. These events keep the business afloat. That is, *Why* is the business in business?	Order, Return, Complaint, Withdrawal, Deposit, Compliment, Inquiry, Trade, Claim
How	Documentation of the event of interest to the enterprise. Documents provide the evidence that the events occurred, such as a Purchase Order recording an Order event. That is, *How* do we know that an event occurred?	Invoice, Contract, Agreement, Account, Purchase Order, Speeding Ticket, Packing Slip, Trade Confirmation
Measurement	Counts, sums, etc. of the other categories (what, where) at or over points in time (when).	Sales, Item Count, Payments, Balance

Table 7 Commonly Used Entity Categories

An **entity instance** is a particular occurrence of an **entity**.

Usage	Entity	Entity Type	Entity Instance
Common Use	Jane	Employee	
Recommended Use	Employee		Jane

Table 8 Entity, Entity Type, and Entity Instance

Entites are represented graphically by **rectangles.**

- **Definition of entites**: (important as they are core Metadata). Definitions clarify the meaning of business vocabulary, and provide rigor to the business rules governing entity relationships
  - **Clarity:** Easy to read and grasp
  - **Accuracy:** Precise and correct description of the entity
  - **Completeness**: All parts of the definition are present. The scope of uniqueness is in the definition. e.g., examples of code values for a code entity

1.3.3.2 Relationship

An association between entities.

1.3.3.2.1 Relationship Aliases

Other aliases based on scheme:

- **Dimensional**: Navigation path
- **NoSQL**: Edge (Graph Database) or Link (Data Vault)
- **Relational on the physical level**: Constraint or Reference

1.3.3.2.2 Graphical Representation of Relationships

Graphically represented as lines on the diagram

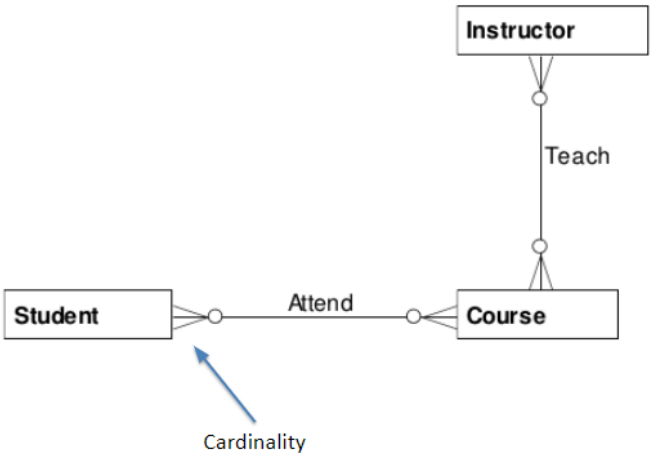
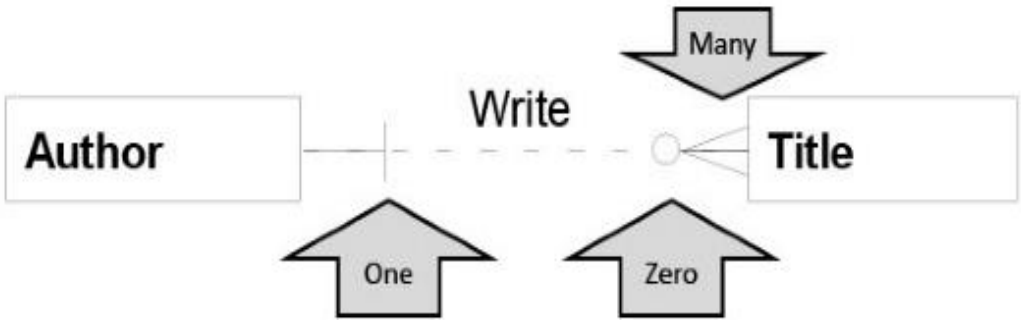


Figure 30 Relationships

Reading a Data Model: a 90 second video by Steve Hoberman
https://www.youtube.com/watch?v=adYohKb47f8



Cardinality Symbols (Steve Hoberman)

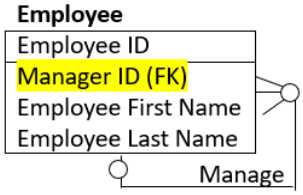
1.3.3.2.3 Relationship cardinality:

Data rules on how entities are connected are enforced. An instance of an entity may have a relationship with zero, one or many instances of another entity.

1.3.3.2.4 Arity of Relationships

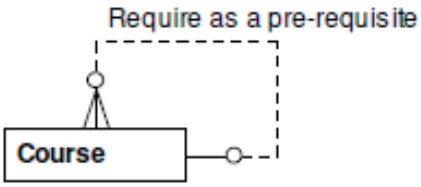
Unary (Recursive or self-referencing) relationship:

EMPLOYEE_ID	FIRST_NAME	LAST_NAME	MANAGER_ID
107	Diana	Lorentz	103
103	Alexander	Hunold	102
102	Lex	De Haan	100
100	Steven	King	<null>



This is a nonidentifying, nonmandatory relationship in which the same entity is both the parent and the child. There is only one entity. **A Foreign Key is used to distinguish between the roles or instances**.

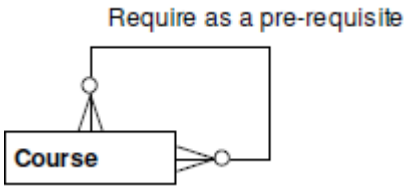
Unary Relationship Explained (Author Unknown)



Hierarchy: One to many recursive relationship Child entity has at most 1 parent

A course can be a prerequisite once

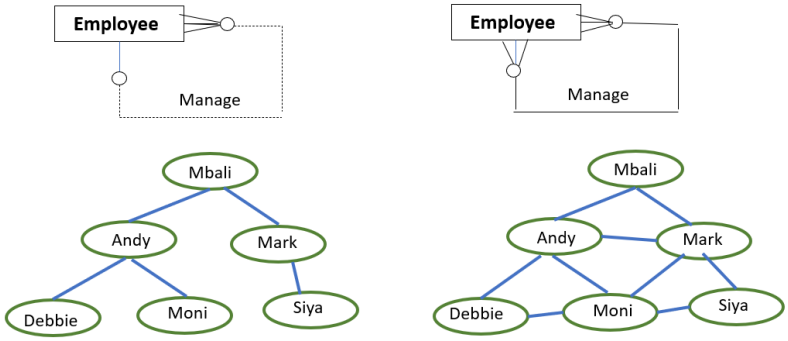
Figure 32 Unary Relationship - Hierarchy



Network: Many to many recursive relationship. Entity instance can have more than one parent.

A course can be prerequisite for many courses

Figure 33 Unary Relationship - Network



Unary Relationship Expanded

Binary Relationship: 2 Entities

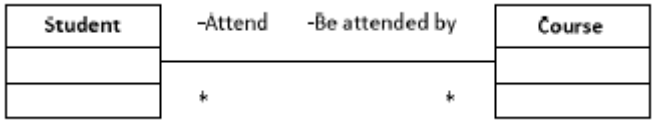


Figure 34 Binary Relationship

Ternary Relationship: 3 Entities.

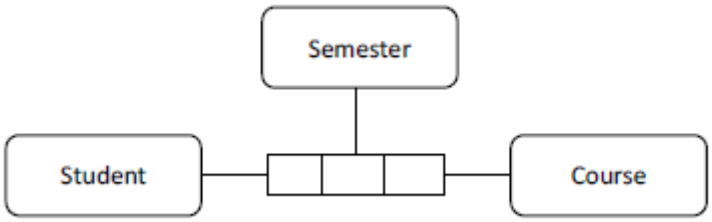


Figure 34 Ternary Relationship

Used in **Fact Based Modelling** for ontologies.

1.3.3.2.5 Foreign Keys

Used in logical and physical data models to represent a relationship.

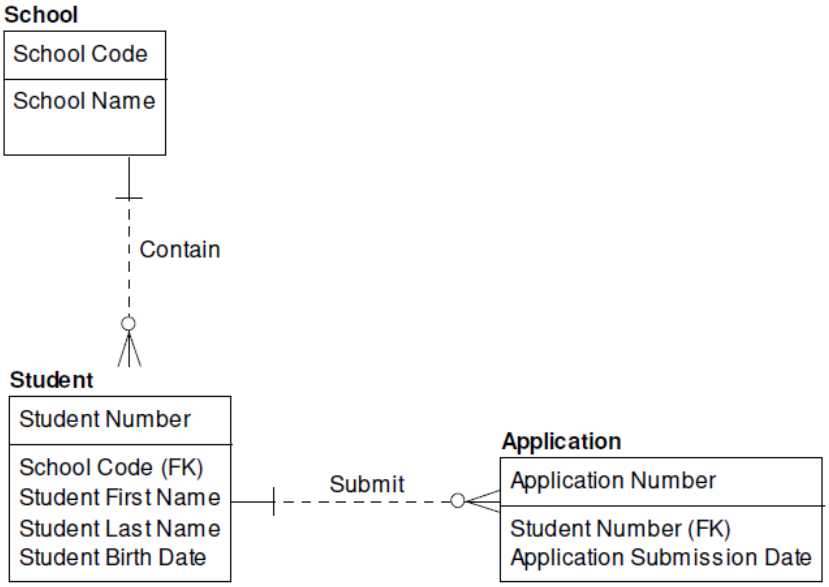


Figure 48 (Used here to demonstrate Foreign Keys in a Non-Identifying Relationship)

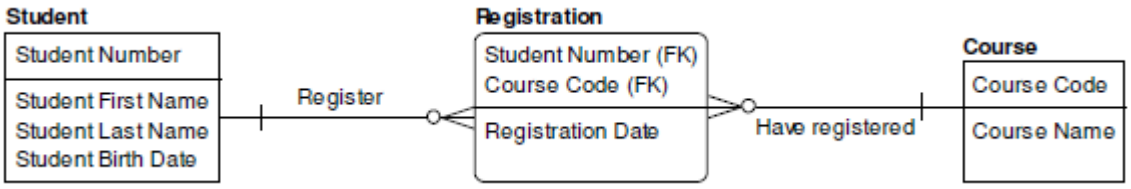


Figure 36 Foreign Keys (Used here to demonstrate Foreign Keys in an Identifying Relationship)

1.3.3.3 Attribute

- A property that identifies, describes or measures an entity
- A column in a table. May have domains
- In a list as in Student above

1.3.3.3.1 Graphical Representation of Attributes

A list within the entity rectangle

1.3.3.3.2 Identifiers (Key)

Set of one or more attributes that uniquely identifies and instance of an entity.

Types of keys (by construction and function):

1.3.3.3.2.1 Construction-type keys:

- **Simple**: One attribute e.g., VIN number or Universal Product Code (UPC)
- **Surrogate:** Unique identifier, often a counter and always system generated, without intelligence and not visible to end users. Maintain uniqueness, navigation, and facilitate integration across applications
- **Composite:** A candidate key that consists of two or more attributes that together uniquely identify an entity
- **Compound:** A compound key is a composite key for which each attribute that makes up the key is a foreign key in its own right

1.3.3.3.2.2 Function type keys:

- **Super Key**: any set of attributes that uniquely identify an instance
- **Candidate key**: Minimal set of attributes (simple or compound) that uniquely identifies an entity instance. Also called **Business** or **Natural** keys
- **Primary key**: The candidate key chosen to be the unique identifier. Usually a **surrogate** key
- **Alternate key**: Unique candidate key, but not chosen for primary. Usually a **business** key. Note: Alternate keys enforce the Data Quality dimension Uniqueness)

1.3.3.3.2.3 Identifying vs Non-Identifying Relationships

Independent Entity:

- Primary key contains only attributes belonging to that entity
- Non-identifying (weak) relationship ---------
- FK non-primary key attribute

Dependent Entity:

- The primary key contains at least one attribute from another entity
- Rectangles with rounded corners
- Identifying (strong) relationship _____

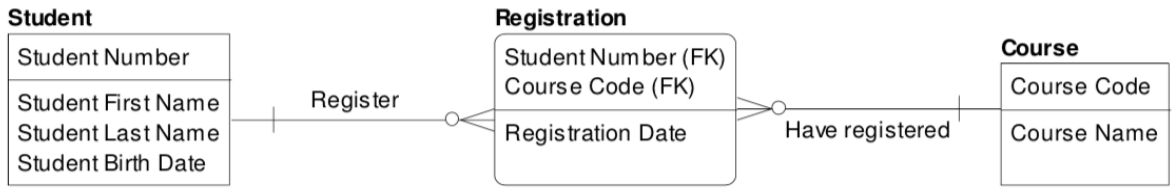


Figure 38 Dependent and Independent Entity

1.3.3.4 Domain

A domain is a complete set of possible values that an attribute can be assigned. It is used to standardise the characteristics of the attributes.

Domains can be restricted by adding **constraints** (additional business rules). Constraints can relate format and/or logic.

Defining domains:

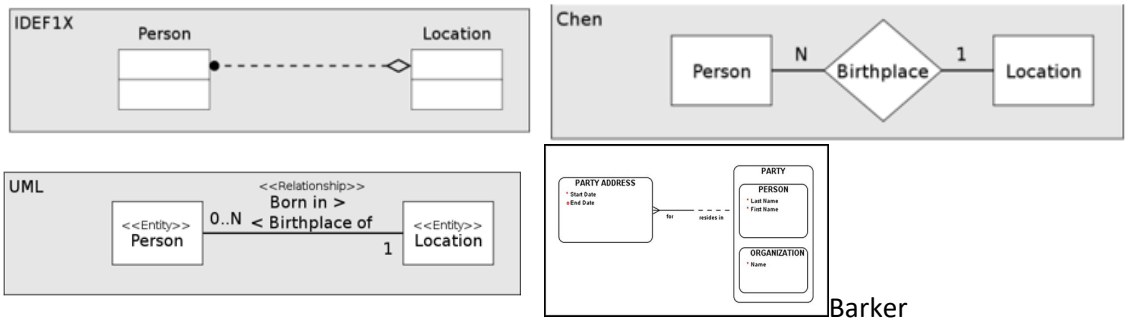
- **Data Type:** Standard type of data in an attribute assigned to that domain, e.g. Integer, Date
- **Data Format:** Domains that use patterns, templates, masks and character limitations
- **List:** Finite set of values
- **Range:** Values of same data type between max and min values or can be open ended

- **Rule-based**: e.g., Compare values to a calculated value

1.3.4 Data Modelling Schemes and Notations

Scheme	Sample Notations
Relational	Information Engineering (IE) Integration Definition for Information Modeling (IDEF1X) Barker Notation Chen
Dimensional	Dimensional
Object-Oriented	Unified Modeling Language (UML)
Fact-Based	Object Role Modeling (ORM or ORM2) Fully Communication Oriented Modeling (FCO-IM)
Time-Based	Data Vault Anchor Modeling
NoSQL	Document Column Graph Key-Value

Table 9 Modelling Schemes and Notations



Barker

Examples of Modelling Notations (Author Unknown)

Scheme	Relational Database Management System (RDBMS)	Multidimensional Database Management System (MDBMS)	Object Databases	Document	Column	Graph	Key-Value
Relational	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Dimensional	CDM LDM PDM	CDM LDM PDM					
Object-Oriented	CDM LDM PDM		CDM LDM PDM				
Fact-Based	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Time-Based	PDM						
NoSQL			PDM	PDM	PDM	PDM	PDM

Table 10 Scheme to Database Cross Reference

The models which can be built for each scheme, based on the technology.

1.3.4.1 Relational

Dr Edgar F. Codd (1970) proposed that data could be managed most effectively in terms of two dimensional relations, reducing redundancy and data storage. This approach was based on the mathematics of set theory. (Note: DMBOK erroneously refers to Dr Codd as Edward)

Relationships capture business rules.

Exact expression of business data, and have one fact in one place (removal of redundancy). Ideal for operational systems requiring quick data entry, individual transaction processing and accurate storage.

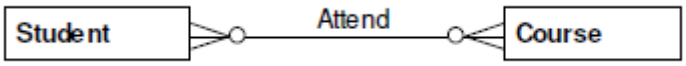
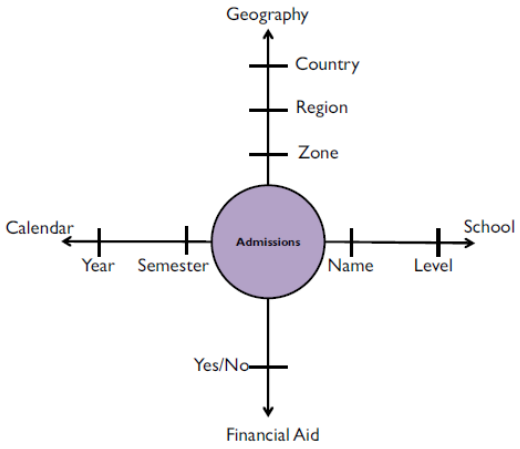


Figure 39 Information Engineering (IE) Notation

1.3.4.2 Dimensional

General Mills and Dartmouth College in the 1960s. Data is structured to optimise query and analysis of large amounts of data. Batch processing.

Dimensional models capture business questions focussed on a particular business process. Relationships capture navigation paths to answer the business question.



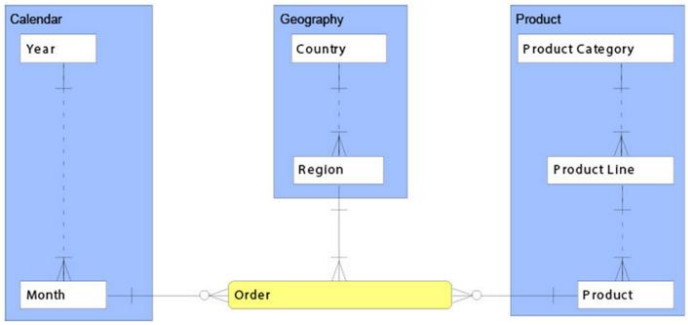
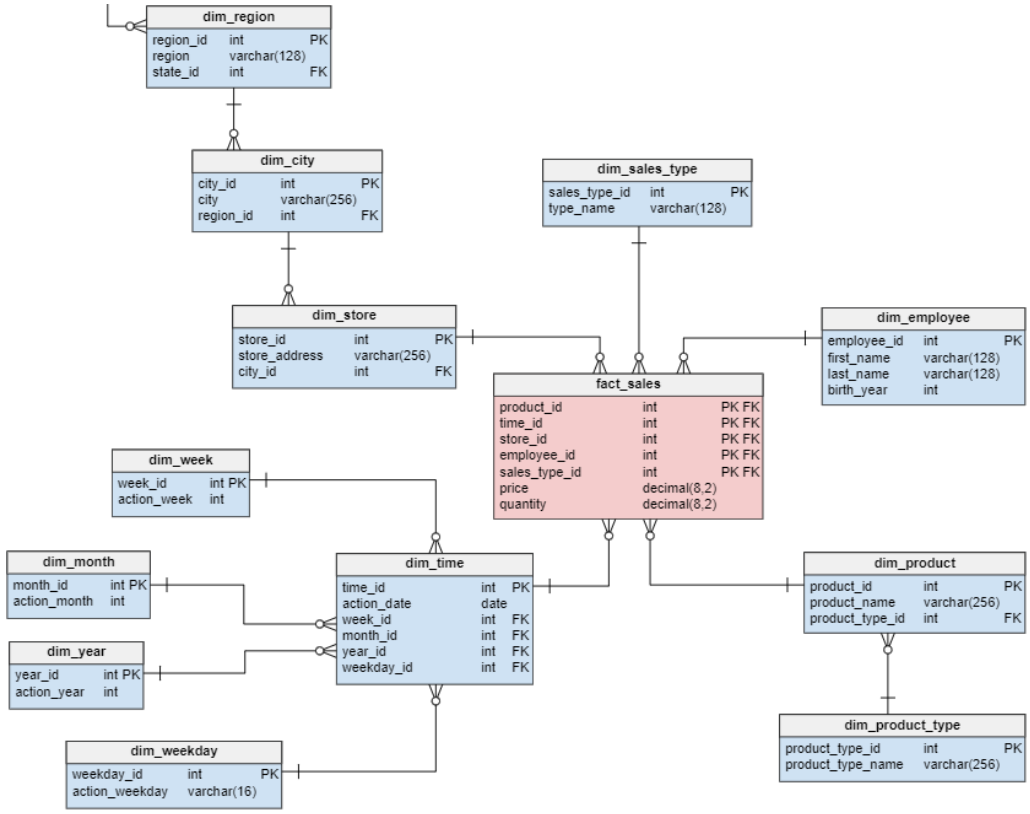
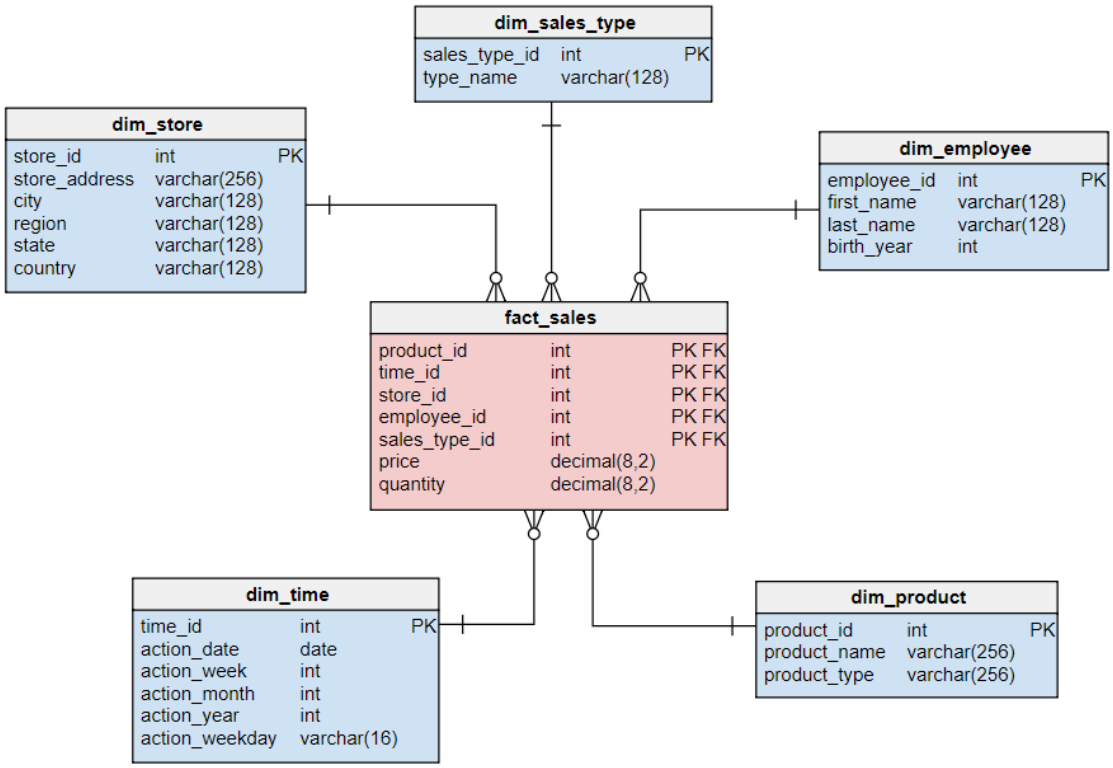


Figure 40 Axis Notation for Dimensional Models Dimensional Conceptual Model Notation

- **Fact Tables:**
  - The measurement, result of calculations or algorithms
  - Numeric
  - Metadata is critical for proper understanding and usage
  - Fact tables consist of a **large number of rows** and few columns
  - Takes up approximately 90% of space in dimensional databases
- **Dimension Tables:**
  - Mostly textual
  - Used for querying
  - Highly denormalised (Note: Correct term is flattened. Normalisation is a Relational term)
  - Must have a unique identifier for each row
  - **Slowly Changing Dimensions** (SCDs) manage changes of attributes based on the rate and type of change:
    - **Overwrite (Type 1):** New value overwrites the old value in place
    - **New Row (Type 2):** New values written a new row, old row marked as not current
    - **New Column (Type 3):** Multiple instances of a value are listed in columns in the same row
- **Snowflaking:** Normalising the star schema (Note: This process is not relational normalisation, but rather the separation of levels in the dimension)
- **Grain:** The most detail a row in the fact table can have
- **Conformed Dimensions:** Built with the entire organisation in mind and shared across dimensional models
- **Conformed Facts:** Use standardised definitions of terms across marts



Example of a Dimensional Logical Data Model Showing Snowflaking (Author Unknown)



Example of a Dimensional Physical Data Model Showing the Flattening of the Dimensions (Author Unknown)

1.3.4.3 Object Oriented (UML)

Unified Modelling Language (UML) class model is used for object oriented databases. Specifies classes (entity types) and their relationship types.

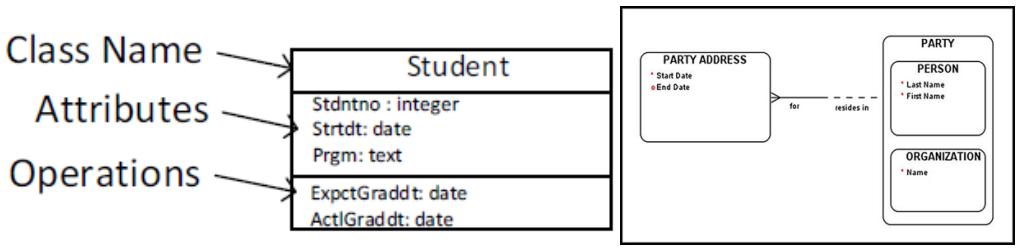


Figure 41 UML Class Model

1.3.4.4 Fact based Modelling

Based on analysis of natural verbalisation in the business domain:

- **Object role modelling (ORM):** Model driven engineering approach which verbalises required information at a conceptual level in a controlled and unambiguous natural language

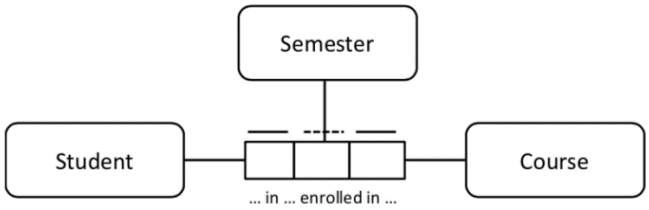


Figure 42 ORM model

- **Fully communication oriented modelling (FCO-IM):**

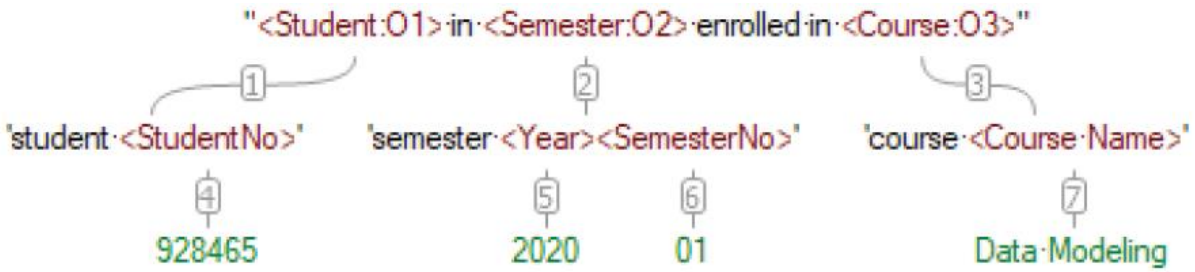


Figure 43 FCO-IM Model

1.3.4.5 Time-Based

Time-based patterns are used when data values must be in chronological order and with time values. Used for data warehouses in a RDBMS environment.

- **Data Vault**: Detail oriented, time-based, normalised tables that support one or more functional areas of the business. Between 3NF and Star schema to meet needs of enterprise data warehouses
    - **3 types of entities**:
        - **Hubs**: the primary key
        - **Links**: provide transaction integration between hubs
        - **Satellites**: provide the context of the hub primary key

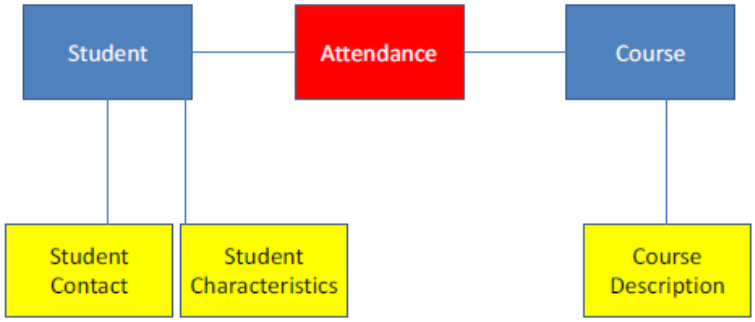


Figure 44 Data Vault Model

- **Anchor Modelling**: Graphical notation for information that changes over time in both structure and content. Similar to traditional data modelling but with extensions for temporal data.
    - **4 basic concepts**:
        - **Anchors**: model entities and events
        - **Attributes**: model properties of anchors
        - **Ties**: model the relationship between anchors
        - **Knots**: model shared properties such as states

On the anchor model in Figure 45, **Student**, **Course**, and **Attendance** are anchors, the gray diamonds represent ties, and the circles represent attributes.

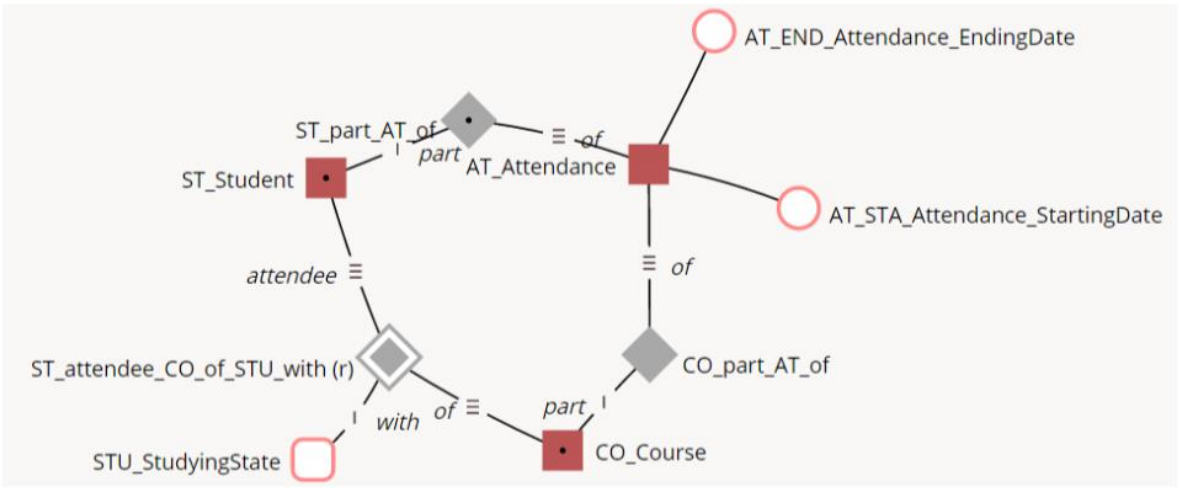


Figure 45 Anchor Model

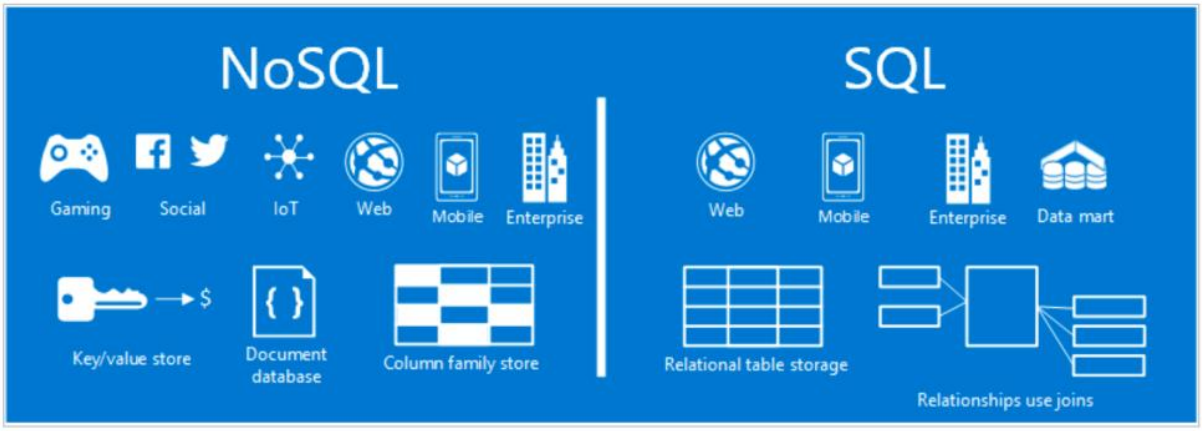
1.3.4.6 NoSQL

NoSQL is the name for databases built on non-relational technology. There are four types of NoSQL databases:

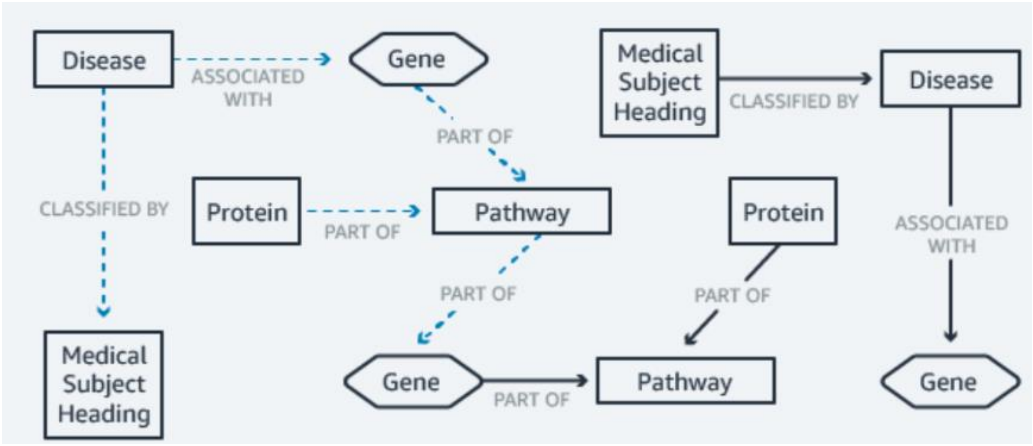
- **Document:**
    - Store the business subject in one structure called a document
    - For example, instead of storing Student, Course, and Registration information in three distinct relational structures, properties from all three will exist in a **single document** called Registration
- **Key-value:**

- Key-value databases allow an application to store its data in only two columns ('key' and 'value')
- Value column can store anything (such as simple information, unformatted text, images, audio, and video**)**

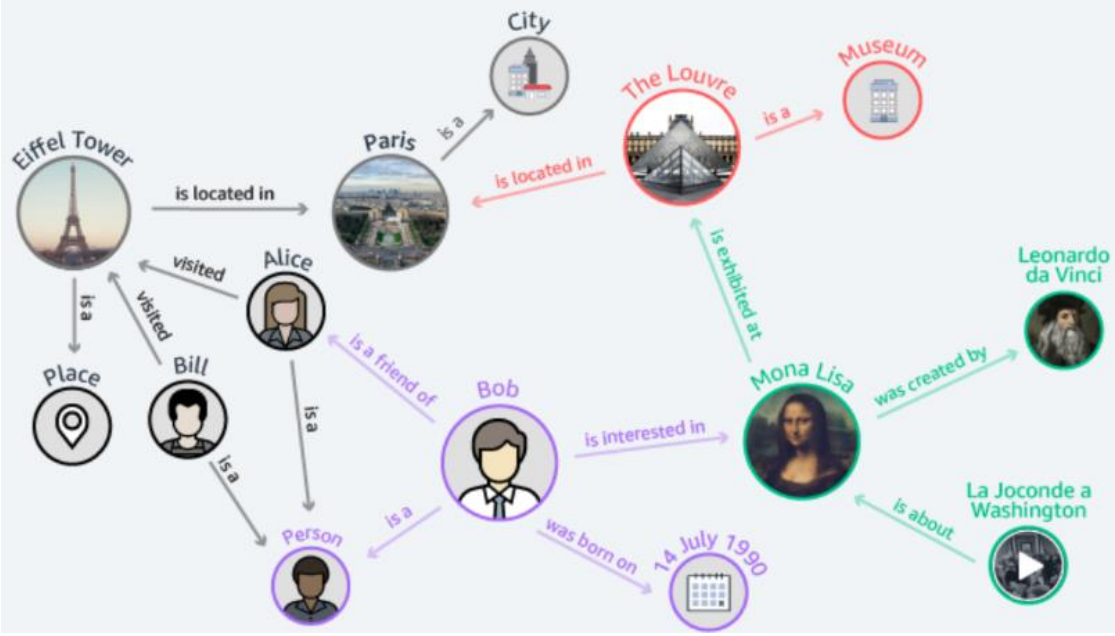
- **Column-oriented:**
  - RDBMSs work with a predefined structure and simple data types, such as amounts and dates, whereas column-oriented databases, such as Cassandra, can work with more complex data types including unformatted text and imagery
  - Store each column in its own structure
- **Graph**:
  - A graph database is designed for data whose relations are well represented as a set of nodes with an undetermined number of connections between these nodes



Sources and Structures for NoSQL and SQL databases (Author Unknown)



Knowledge Graph Example (Model) (Author Unknown)



Knowledge Graph Example (Data) (Author Unknown)

1.3.5 Data Model Levels of Detail

- **Conceptual** – "Real world" view of the enterprise modelled in the database
- **External** – Subsets of the enterprise model which represent the data requirements in a specific usage context. Commonly translates to Logical Model
- **Internal** – Internal schema or machine view. Describes the stored representation of the enterprise's data. Commonly translates to the Physical Model

1.3.5.1 Conceptual (CDM)

The conceptual data model captures the high-level data requirements as a collection of related concepts. Basic business entities and the relationships between them (business rules) are described.

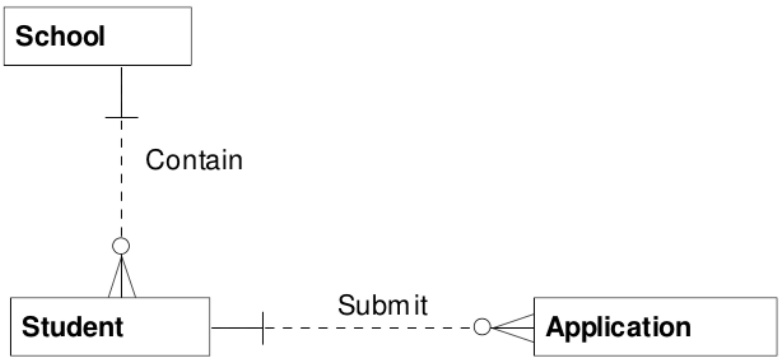


Figure 46 Relational Conceptual Model

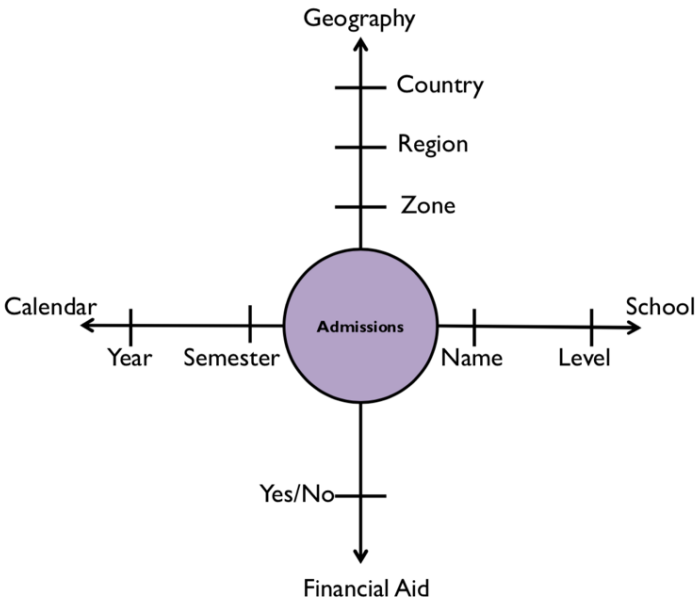


Figure 47 Dimensional Conceptual Model

1.3.5.2 Logical (LDM)

A detailed, technology independent representation in a specific usage context. An extension of conceptual model. Attributes are assigned by applying normalisation.

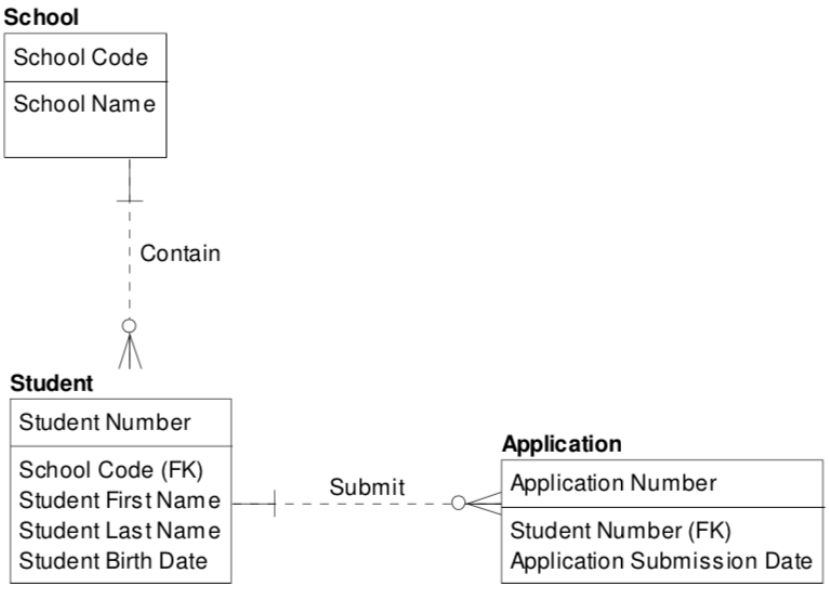


Figure 48 Relational Logical Data Model

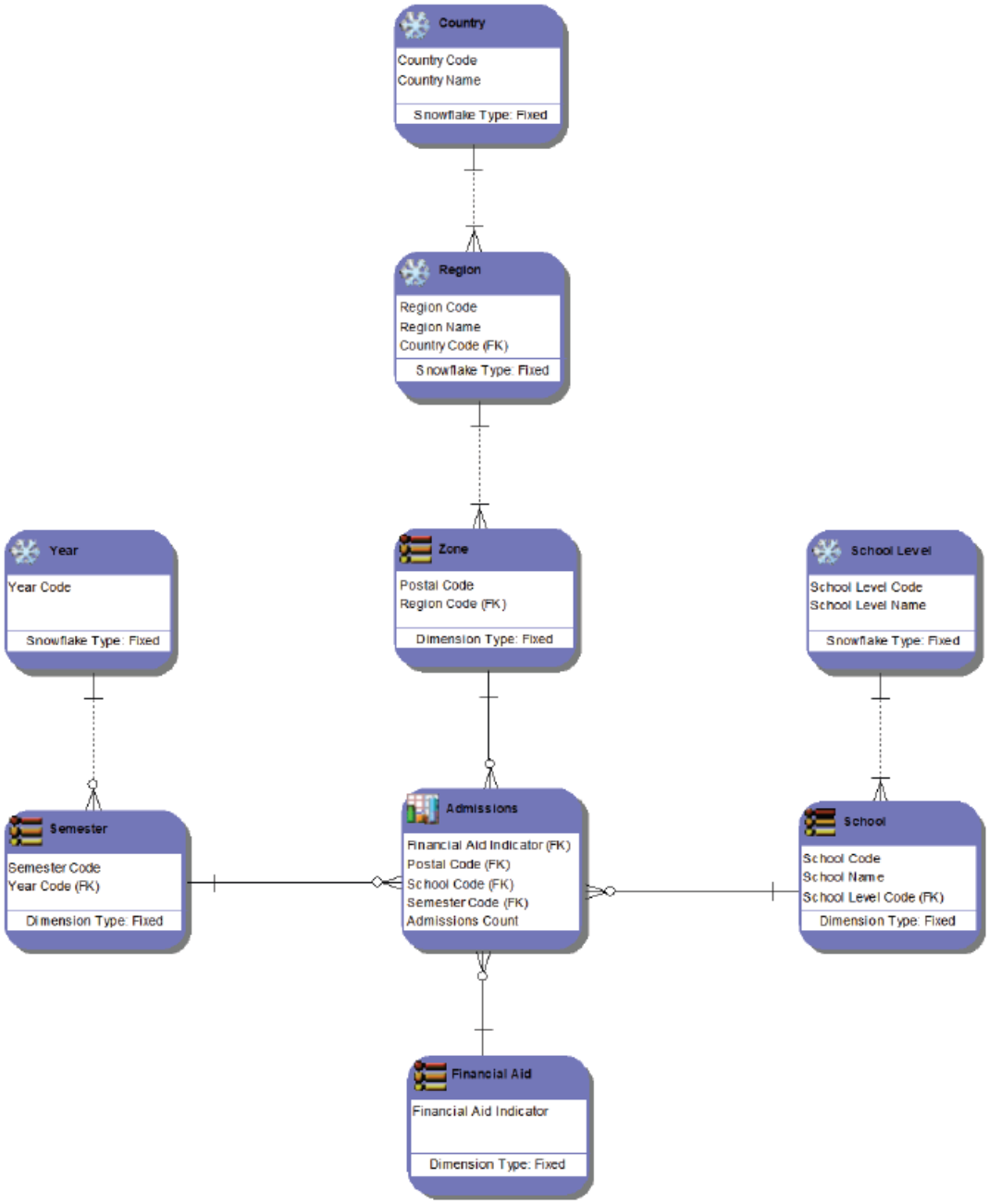


Figure 49 Dimensional Logical Data Model

1.3.5.3 Physical (PDM)

Detailed technical solution using the logical data model. Built for a particular technology of DBMS.

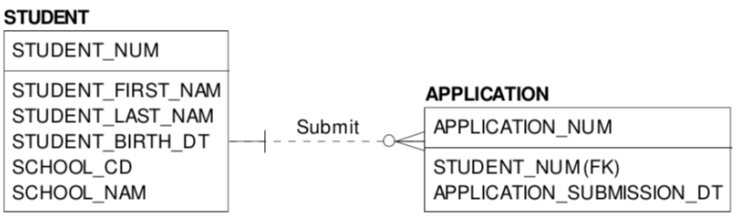


Figure 50 Relational Physical Data Model

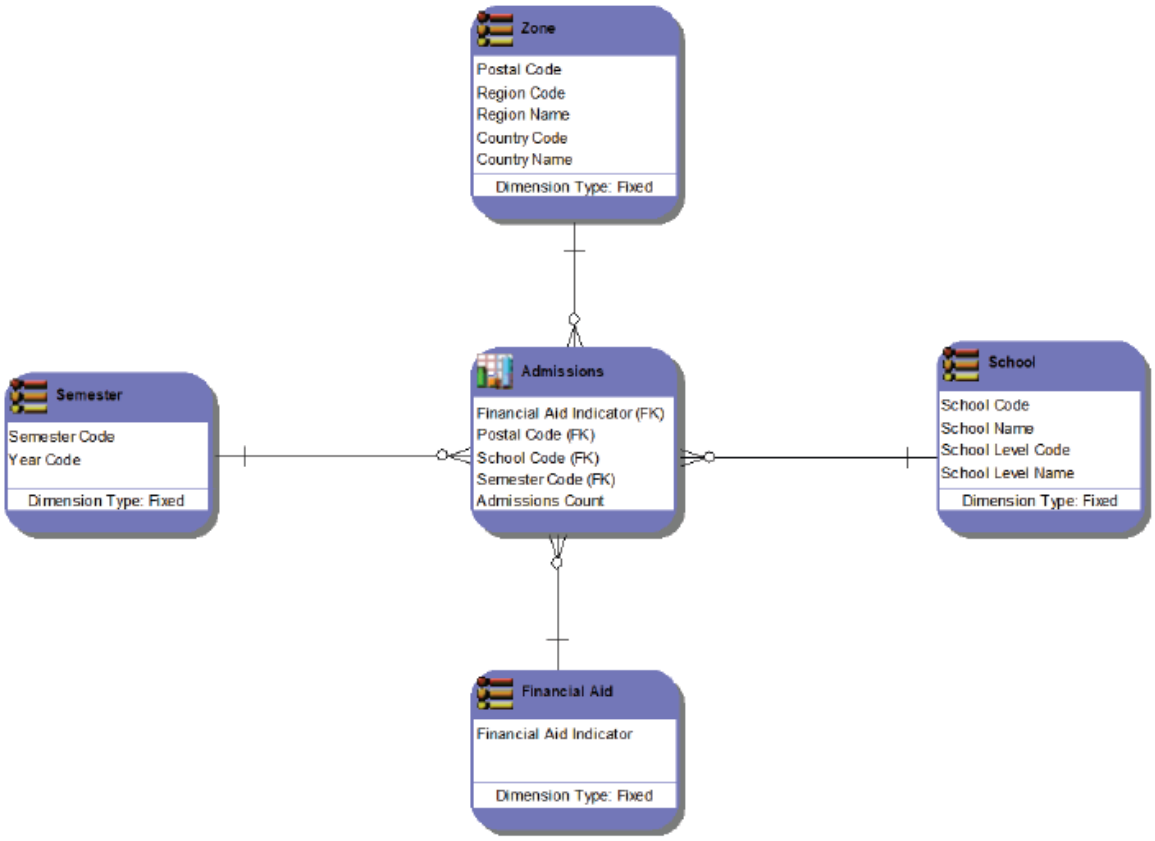


Figure 50 Dimensional Physical Data Model

1.3.5.3.1 Canonical

A **Canonical Model** is a type of physical data model which describes the structure of data moving between systems, the structure the sender and receiver should use. Structures should be generic and reusable. Used when sending data through web services, the Enterprise Service Bus (ESB) or Enterprise Application Integration (EAI). May be instantiated on middleware.

1.3.5.3.2 Views

Virtual tables used to simplify data access, control data access and rename columns without redundancy and loss of referential integrity due to denormalisation.

1.3.5.3.3 Partitioning

Process of splitting a table to improve performance

- **Vertically split**: subset tables contain subsets of columns
- **Horizontally split**: a value in a column is a delimiter to create subset tables

1.3.5.3.4 Denormalisation

- Deliberate transformation of normalised logical model entities into physical tables with redundant data structures
- Done to improve performance
- Called Embedding in Document databases, and Collapsing or Combining in Dimensional databases.
- Can introduce Data Quality issues due to duplication. Implement DQ checks

1.3.6 Normalisation

Normalisation is the process of applying rules to organise business complexity into stable data structures. The basic goal is to eliminate redundancy by keeping each attribute in only one place. Requires understanding of attribute's relationship to its primary key.

Normalisation levels:

- **First normal form (1NF)**: Valid primary key. Each attribute depends on key. Eliminate repeating groups and ensure each attribute is atomic and not multi valued.
- **Second normal form (2NF)**: Each entity has the minimal primary key and each attribute depends on the complete primary key
- **Third normal form (3NF)**: no hidden primary keys. Each attribute depends on no attributes outside the key ("the key, the whole key and nothing but the key")
- **Boyce / Codd normal form (BCNF)**: Resolves overlapping composite candidate keys (hidden business rules)
- **Fourth normal form (4NF)**: Resolves all many to many relationships in pairs until they can't be broken down into smaller pieces
- **Fifth normal form (5NF)**: Resolves inter-entity dependencies into basic pairs, and all join dependencies use parts of primary keys

See end of chapter for the example slides.

1.3.7 Abstraction

Abstraction is the removal of details in such a way as to broaden applicability to a wide class of situations while preserving the important properties, e.g., Party/Role structure.

Includes:

- **Generalisation**: Groups common attributes and relationships into **supertype** entities
- **Specialisation**: Separates distinguishing attributes within an entity into **subtype** entities
    - Subtypes can also be created using **roles** or **classification** e.g., Party with subtypes Individual and Organisation

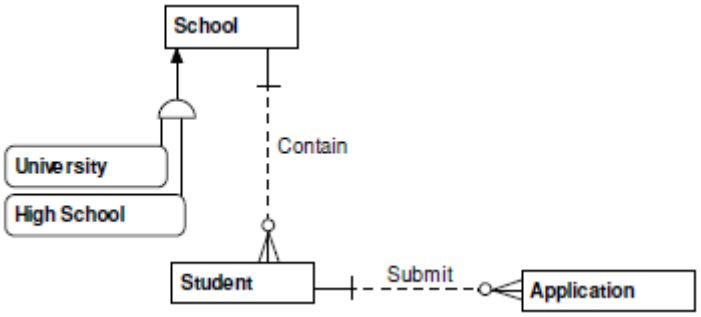


Figure 52 Supertype and Subtype Relationships

2 Activities

2.1 Plan for Data Modelling

Tasks to plan for:

- Evaluating organisational requirement
- Creating standards

- Determining data model storage

Deliverables of the modelling process:

- **Diagram:** The visual that captures the requirements in precise form. Depicts:
  - **Level of detail:** Conceptual, logical or physical
  - **Scheme:** Relational, Dimensional, Object-oriented, Fact-based, Time-based or NoSQL
  - **Notation:** information engineering, unified modelling language, object role modelling
- **Definitions** for entities, attributes and relationships are essential for precision
- **Issues and outstanding questions:** Document current issues
- **Lineage**: Where the data comes from
  - A source/target mapping: Can see source system attributes and how they populate target system attributes
  - Trace components from conceptual to logical to physical
  - Data modeller obtains a good understanding of data requirements and can determine the source attributes
  - Source attributes can be used to check accuracy of the model and mapping

2.2 Build the Data Model

Modelling involves studying previous analysis, existing data models and databases etc.

Modelling is an iterative process: Draft the model then go back to business analysts for clarification, update the model then ask more questions. This increases the precision of the model.

2.2.1 Forward Engineering

Build a new application beginning with the requirements.

- CDM: Understand the scope and key terminology
- LDM: Documents the business solution
- PDM: Documents the technical solution

2.2.1.1 Conceptual Data Modelling

Steps to create a CDM:

- **Select scheme:** Relational, dimensional, fact-based or NoSQL
- **Select notation:** Appropriate notation for selected scheme
- **Complete initial CDM:** Capture the viewpoint of the user group
  - Collect the highest-level concepts (nouns)
  - Collect the activities (verbs) that connect these concepts. Relationships can go both ways or involve more than two concepts
- **Incorporate enterprise terminology:** Ensure consistency with enterprise terminology and rules
- **Obtain signoff:** Reviewed for data modelling best practices, and that it meets requirements

2.2.1.2 Logical Data Modelling

LDM captures the data requirements within the scope of the CDM.

- **Analyse information requirements:** Elicitation, organisation, documentation, review, refinement approval and change control of business requirements
- **Analyse existing documentation:** Pre-existing artefacts provide a jump start for new model

- **Add associative entities:** Used to describe the many-to-many relationships
- **Add attributes:** Should be atomic
- **Assign domains:** Allow for consistency of value sets and format
- **Assign keys:** Identify primary and alternate keys
- **Normalise**

2.2.1.3 Physical Data Modelling

LDM must be modified to perform well within storage applications.

- **Resolve logical abstractions:** Subtypes and supertypes become separate entities
- **Add attribute details:** Technical names, physical domain, physical data types and length of fields as well as constraints such as NOT NULL
- **Add Reference Data objects:** Small Reference Data value sets
- **Assign surrogate keys:** Unique key values not visible to business
- **Denormalise for performance:** The performance benefit of adding redundancy outweighs the cost
- **Index for performance:** optimise query performance. Prevents every row being read (table scan)
- **Partition for performance:** Partition on a date key is usually recommended
- **Create views:** Requirements driven. Control access to certain data elements or embed joins or filters

2.2.2 Reverse Engineering

The process of documenting an existing database.

- PDM first to understand the technical design
- LDM to document the business solution met by the system
- CDM documents the scope and key terminology

2.3 Review the Data Model

Apply a data quality verifier such as Steve Hoberman's Data Model Scorecard ® for quality control.

2.4 Maintain the Data Models

The data models need to be kept current. Update when business requirements or processes change. Compare physical with logical regularly.

3 Tools

- **Data Modelling Tools:**
  - May support forward engineering from conceptual to logical to physical including DDL generation
  - **Rubber-banding** is the automatic redrawing of relationship lines when an entity is moved
  - Most support reverse engineering from database to CDM
  - Some support naming standards and metadata storage
- **Lineage Tools:**
  - Capture and maintenance of source structures for each attribute in the data model
  - Excel is most commonly used
- **Data Profiling Tools:**

    - Explores data content and validates it according to metadata and identifies data quality gaps/deficiencies
- **Metadata Repositories:**
    - Store descriptive data about the model, including diagram and definitions
    - Enable sharing
- **Data Model Patterns:**
    - Reusable modelling structures
- **Industry Data Models:**
    - Pre-built for an entire industry
    - Needs to be customised

4 Best Practices

4.1 Best Practices in Naming Conventions

- ISO 11179 Metadata Registry is an international standard for representing Metadata, consists of sections related to data standards such as naming attributes and definitions
- Data architects, data analysts and database administrators jointly develop standards for an organisation, to complement related IT standards
- Names should be unique and descriptive
- Logical names must be meaningful to business users
- Physical names must conform to DBMS restrictions

4.2 Best Practices in Database Design

Design principles for the DBA (PRISM):

- **Performance and ease of use:** Quick and easy access
- **Reusability:** Multiple applications can use the data
- **Integrity:** Data should always have valid business meaning and value
- **Security:** Only available to authorised users
- **Maintainability:** Ensure data work yields value, and the cost of creating, storing, maintaining, using and disposing of data does not exceed its value to the organisation

5 Data Model Governance

5.1 Data Model and Design Quality Management

Data models and database designs should be a reasonable balance between the short term needs and the long term needs of the enterprise.

5.1.1 Develop Data Modelling and Design Standards

Data modelling and database design standards help meet business data requirements, conform to Enterprise and Data Architecture, and ensure data quality. Standards should include the following:

- List and description of standards data modelling and database design deliverables
- List of standard names, abbreviations and abbreviation rules
- List of standard naming formats for all data model objects
- List and description of standard methods of creating and maintaining these objects
- List and description of data modelling and database design roles and responsibilities
- List and description of all Metadata properties captured in data modelling and database design

- Metadata quality expectations and requirements
- Too use guidelines
- Guidelines for preparing and leading design reviews
- Guidelines for versioning models
- Practices that are discouraged

5.1.2 Review Data Model and Database Design Quality

- **Requirements Reviews:** Project team conducts this review
  - Starting model if there is one
  - Changes made to the model
  - Options considered and rejected
  - How well new model conforms to modelling and architectural standards
- **Design reviews:** Group of diverse subject matter experts
  - Chair meeting with an agenda to maintain order and move forward
  - Participants are given required documentation and chair solicits input
  - Summarise group's consensus finding
  - If there is no approval:
    - Modeller reworks
    - Final say should be given by the owner of the system

5.1.3 Manage Data Model Versioning and Integration

Preserve the lineage of changes:

- **Why** the project or situation required the change
- **What** and **how** the object changed
- **When** the change was approved and made to the model
- **Who** made the change
- **Where** the change was made (which model)

Modelling tools may have a repository which provides versioning and integration, else preserve models on DDL exports or XML files in a standard source code management system.

5.2 Data Modelling Metrics

Steve Hoberman's Data Model Scorecard® provides a way to measure the quality of a data model and identifies specific areas for improvement.

#	Category	Total score	Model score	%	Comments
1	How well does the model capture the requirements?	15			
2	How complete is the model?	15			
3	How well does the model match its scheme?	10			
4	How structurally sound is the model?	15			
5	How well does the model leverage generic structures?	10			
6	How well does the model follow naming standards?	5			
7	How well has the model been arranged for readability?	5			
8	How good are the definitions?	10			
9	How consistent is the model with the enterprise?	5			
10	How well does the metadata match the data?	10			
	TOTAL SCORE	100			

Table 11 Data Model Scorecard® Template

Description of each category:

1. **How well does the model capture the requirements?** The model supports all required queries for example.
2. **How complete is the model?** Completeness of requirements and completeness of Metadata, nothing extra, nothing missing.
3. **How well does the model match its scheme?** Level of detail (conceptual, logical or physical) and the scheme (Relational, dimensional, NoSQL etc.) matches the definition of the type of model being reviewed.
4. **How structurally sound is the model?** Validate the design practices to ensure a database can be built.
5. **How well does the model leverage generic structures?** Appropriate level of abstraction.
6. **How well does the model follow naming standards?** Ensure correct and consistent naming standards have been applied to the model.
7. **How well has the model been arranged for readability?** Parent entities above child, groupings and shorter relationship lines improve readability.
8. **How good are the definitions?** Clear, complete and accurate.
9. **How consistent is the model with the enterprise?** If and Enterprise Data Model exists.
10. **How well does the Metadata match the data?** Confirm the actual data to be stored fits in the model.

5.3 Normalisation Example (adapted from Steve Hoberman's Mastering Data Modelling Masterclass)

Normalisation

- **First normal form (1NF)**: Valid primary key. Each attribute depends on key. Eliminate repeating groups and ensure each attribute is atomic and not multi valued.
- **Second normal form (2NF)**: Each entity has the minimal primary key and each attribute depends on the complete primary key
- **Third normal form (3NF)**: no hidden primary keys. Each attribute depends on no attributes outside the key ("the key, the whole key and nothing but the key")
- **Boyce / Codd normal form (BCNF)** Resolves overlapping composite candidate keys (hidden business rules)
- **Fourth normal form (4NF)** Resolves all many to many relationships in pairs until they can't be broken down into smaller pieces.
- **Fifth normal form (5NF)**: Resolves inter-entity dependencies into basic pairs, and all join dependencies use parts of primary keys.

Chaos Logical Model

Chaos Employee		
P *	Emp ID	Integer
	Dept Cd	String
	Phone 1	String (50)
	Phone 2	String (50)
	Phone 3	String (50)
	Emp Name	String
	Dept Name	String
	Emp Dept Role	String
	Emp role Experience	String
	Emp Staet Date	Date
	Emp Vest Ind	CHAR (1)
Employee_Chaos PK (Emp ID)		

Chaos Employee Data

Emp ID	Dept Cd	Phone 1	Phone 2	Phone 3	Emp Name	Dept Name	Emp Dept Role	Emp Role Experience	Emp Start Date	Emp Vest Ind
1	DG, DM	082-686-9958	011-111-1112,011-111-1113	011-111-1111	Bruce Wayne	Data Governance, Data Modelling	EIM Manager, Data Modeller	5, 10	1 Jan 2018	N
2	DM	082-555-1234	011-111-1113	011-111-1111	Clark Kent	Data Modelling	Data Modeller	6	1 Jan 2018	N
3	MD	084-504-2535	011-111-1114	011-111-1111	Tony Stark	Master Data	Master Data Manager	6	1 Jan 2017	Y

1. Attribute Names are UNCLEAR
    1. Phone 1, Phone 2, Phone 3
2. Data IS NOT single-valued
    1. Dept Cd
    2. Phone 2
    3. Dept-Name
    4. Emp Dept Role
    5. Emp Role Experience

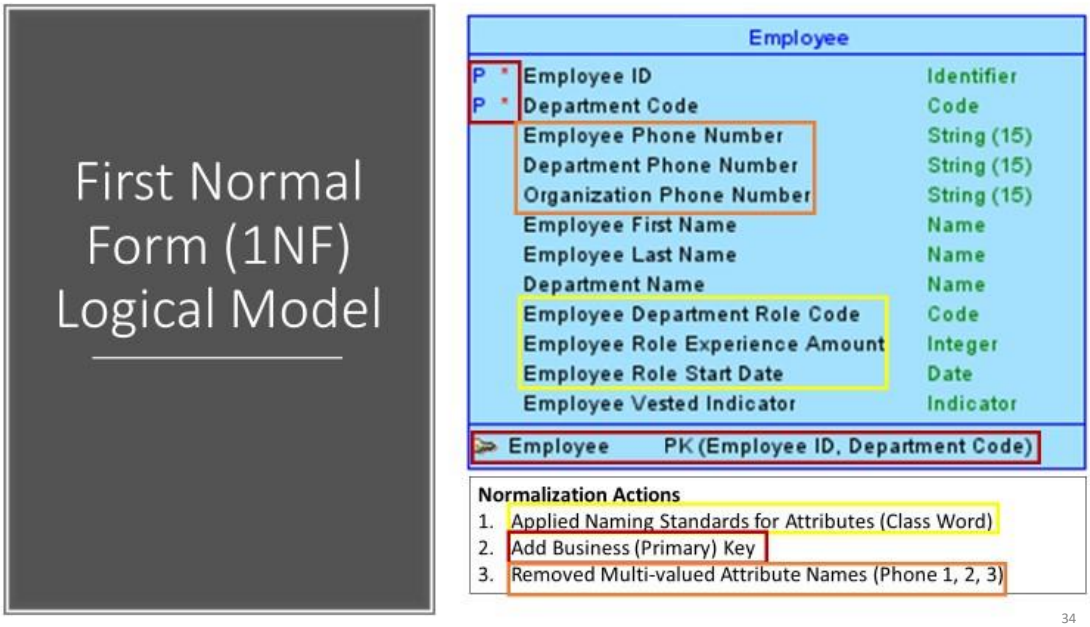
Normalization in a nutshell

Every attribute is single-valued (1NF) and depends completely (2NF) and only on its primary key (3NF).

Steve Hoberman

It is ALL about the Key, the WHOLE Key and NOTHING but the Key

So Help me Codd


First Normal Form (1NF) Logical Model
Employee
P * Employee ID Identifier
P * Department Code Code
Employee Phone Number String (15)
Department Phone Number String (15)
Organization Phone Number String (15)
Employee First Name Name
Employee Last Name Name
Department Name Name
Employee Department Role Code Code
Employee Role Experience Amount Integer
Employee Role Start Date Date
Employee Vested Indicator Indicator
Employee PK (Employee ID, Department Code)
Normalization Actions
1. Applied Naming Standards for Attributes (Class Word)
2. Add Business (Primary) Key
3. Removed Multi-valued Attribute Names (Phone 1, 2, 3)
34

1NF Data Observations

Emp ID	Dept Cd	Employee Phone Number	Department Phone Number	Organisation Phone Number	Emp First Name	Emp Last Name	Dept Name	Emp Dept Role	Emp Role Experience	Emp Start Date	Emp Vest Ind
1	DG	082-686-9958	011-111-1112	011-111-1111	Bruce	Wayne	Data Governance	EIM Manager	5,	1 Jan 2018	N
1	DM	082-686-9958	011-111-1113	011-111-1111	Bruce	Wayne	Data Modelling	Data Modeller	10	1 Jan 2018	N
2	DM	082-555-1234	011-111-1113	011-111-1111	Clark	Kent	Data Modelling	Data Modeller	6	1 Jan 2018	N
3	MD	084-504-2535	011-111-1114	011-111-1111	Tony	Stark	Master Data	Master Data Manager	6	1 Jan 2017	Y

1. Attribute Names are CLEAR
    1. Department Code
    2. Role Experience Amount (Class Word)
2. Attribute Names are SINGLE-VALUED
    1. Phone 1 – Employee Phone Number
    2. Phone 2 - Department Phone Number
    3. Phone 3 – Organization Phone Number
3. Data IS SINGLE-VALUED
    1. No Data Separators

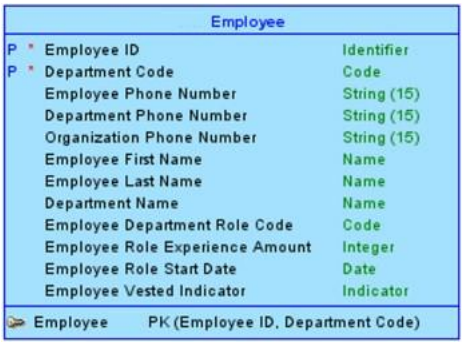
Second Normal Form (2NF)

Ensure minimal set of attributes that uniquely identify each entity instance

- Are all of the attributes in the primary key needed to retrieve a single instance of [[insert attribute name here]]?

Employee

Employee Identifier Department Code
Employee Phone Number Department Phone Number Organization Phone Number Employee First Name Employee Last Name Department Name Employee Start Date Employee Vested Indicator


Employee
P * Employee ID Identifier
P * Department Code Code
Employee Phone Number String (15)
Department Phone Number String (15)
Organization Phone Number String (15)
Employee First Name Name
Employee Last Name Name
Department Name Name
Employee Department Role Code Code
Employee Role Experience Amount Integer
Employee Role Start Date Date
Employee Vested Indicator Indicator
Employee PK (Employee ID, Department Code)

Second Normal Form (2NF) Logical Model – Step 1

2NF Employee
Employee ID Identifier
Employee Phone Number String (15)
Employee First Name Name
Employee Last Name Name
Employee Start Date Date
Employee Vested Indicator Indicator
Employee 2NF PK (Employee ID)
Assign Work
2NF Employee Department Assignment
Employee ID Identifier
Department Code Code
Employee Department Role Name Name
Employee Role Experience Amount Integer
Employee Role Start Date Date
3NF Employee Department PK (Employee ID, Employee Department Role Name, Department Code)
Provide work
2NF Department
Department Code Code
Department Name Name
Department Phone Number String (15)
Organization Phone Number String (15)
2NF Department PK (Department Code)
Department Name AK (Department Name)

Normalization Actions

1. Separated Employee & Department
2. Created a JOIN table to resolve MANY-TO-MANY

2NF Data

	Employee ID	Employee Phone Number	Emp First Name	Emp Last Name	Emp Start Date	Emp Vested Indicator
1	1	082-686-9958	Bruce	Wayne	1 Jan 2018	N
2	2	082-555-1234	Clark	Kent	1 Jan 2018	N
3	3	084-504-2535	Tony	Stark	1 Jan 2017	Y

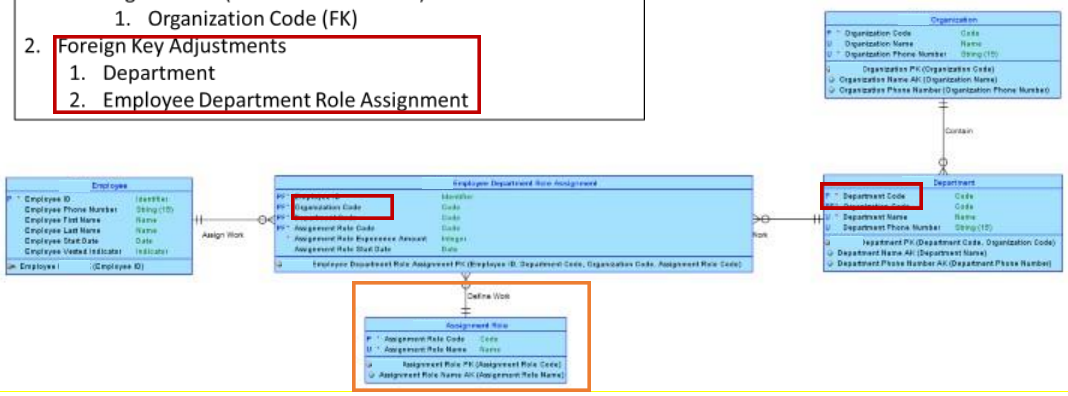
	Employee ID	Department Code	Employee Department Role Name	Employee Role Experience Amount	Employee Start Date
1	1	DG	EIM Manager	5	2018-01-01
2	1	DM	Data Modeller	10	2018-01-01
3	2	DM	Data Modeller	6	2018-01-01
4	3	MD	Master Data Manager	8	2017-01-01

	Department_Code	Department_Name	Department_Phone_Number	Organization_Phone_Number
1	DG	Data Governance	083-111-1112	083-111-1111
2	DM	Data Modelling	083-111-1113	083-111-1111
3	MD	Master Data	083-111-1114	083-111-1111

Second Normal Form- Step 2

Normalization Actions

1. WHOLE KEY
   1. Assignment Role
   2. Organization (No MANY-TO-MANY)
      1. Organization Code (FK)
2. Foreign Key Adjustments
   1. Department
   2. Employee Department Role Assignment



2NF Data: Step 2

	Employee ID	Employee Phone Number	Emp First Name	Emp Last Name	Emp Start Date	Emp Vested Indicator
1	1	082-686-9958	Bruce	Wayne	1 Jan 2018	N
2	2	082-555-1234	Clark	Kent	1 Jan 2018	N
3	3	084-504-2535	Tony	Stark	1 Jan 2017	Y

	Employee ID	Organization Code	Department Code	Assignemnt Role Experience Amount	Assignment Role Start Date
1	1	MDS	DG	5	2019-05-09
2	1	MDS	DM	10	2019-05-09
3	2	MDS	DM	6	2019-05-09
4	3	MDS	MD	8	2019-05-09

	Department_Code	Organization_Code	Department_Name	Department_Phone_Number
1	DG	MDS	Data Governance	083-111-1112
2	DM	MDS	Data Modelling	083-111-1113
3	MD	MDS	Master Data	083-111-1114

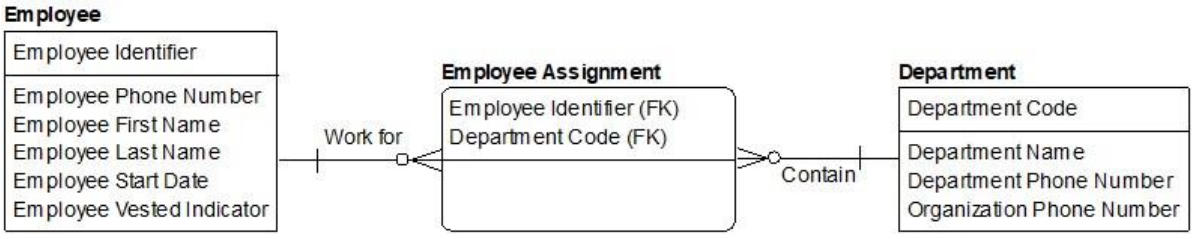
	Organization_Code	Organization_Name	Organization_Phone_Number
1	MDS	Modelware Systems	083-111-1111

	Assignment_Role_Code	Assignment_Role_Name
1	DM	Data Modeller
2	DG	EIM Manager
3	MD	Master Data Manager

Third Normal Form (3NF)

Remove hidden dependencies

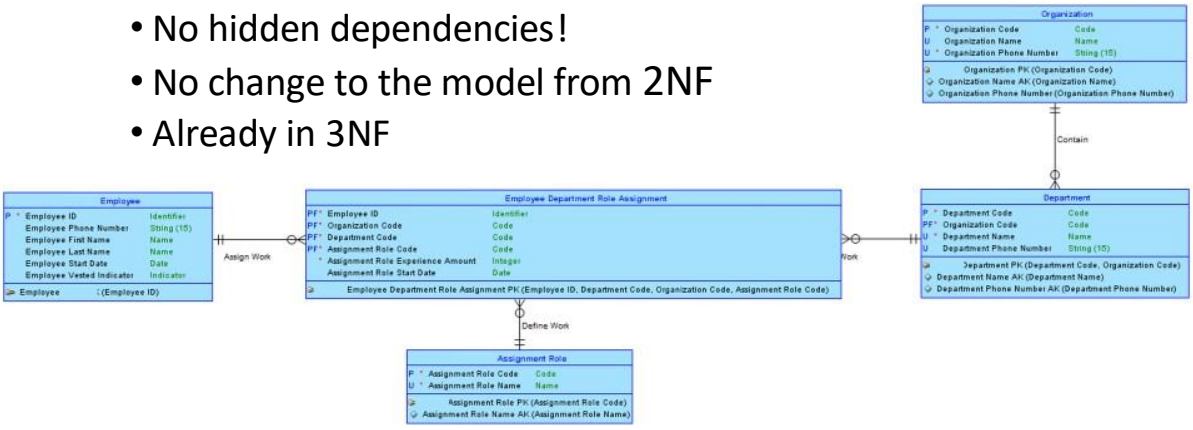
- Is [[insert attribute name here]] a fact about any other attribute in this same entity?



Third Normal Form (3NF)

Remove hidden dependencies

- No hidden dependencies!
- No change to the model from 2NF
- Already in 3NF



Third Normal Form Data

	Employee ID	Employee Phone Number	Emp First Name	Emp Last Name	Emp Start Date	Emp Vested Indicator
1	1	082-686-9958	Bruce	Wayne	1 Jan 2018	N
2	2	082-555-1234	Clark	Kent	1 Jan 2018	N
3	3	084-504-2535	Tony	Stark	1 Jan 2017	Y

	Employee ID	Organization Code	Department Code	Assignment Role Name	Assignemnt Role Experience Amount	Assignment Role Start Date
1	1	MDS	DG	EIM Manager	5	2019-05-09
2	1	MDS	DM	Data Modeller	10	2019-05-09
3	2	MDS	DM	Data Modeller	6	2019-05-09
4	3	MDS	MD	Master Data Manager	8	2019-05-09

	Department_Code	Organization_Code	Department_Name	Department_Phone_Number
1	DG	MDS	Data Governance	083-111-1112
2	DM	MDS	Data Modelling	083-111-1113
3	MD	MDS	Master Data	083-111-1114

	Organization_Code	Organization_Name	Organization_Phone_Number
1	MDS	Modelware Systems	083-111-1111

	Assignment_Role_Code	Assignment_Role_Name
1	DM	Data Modeller
2	DG	EIM Manager
3	MD	Master Data Manager

Chapter 6
Data Storage and Operations

1 Introduction

True custodians of the data – Data at rest.

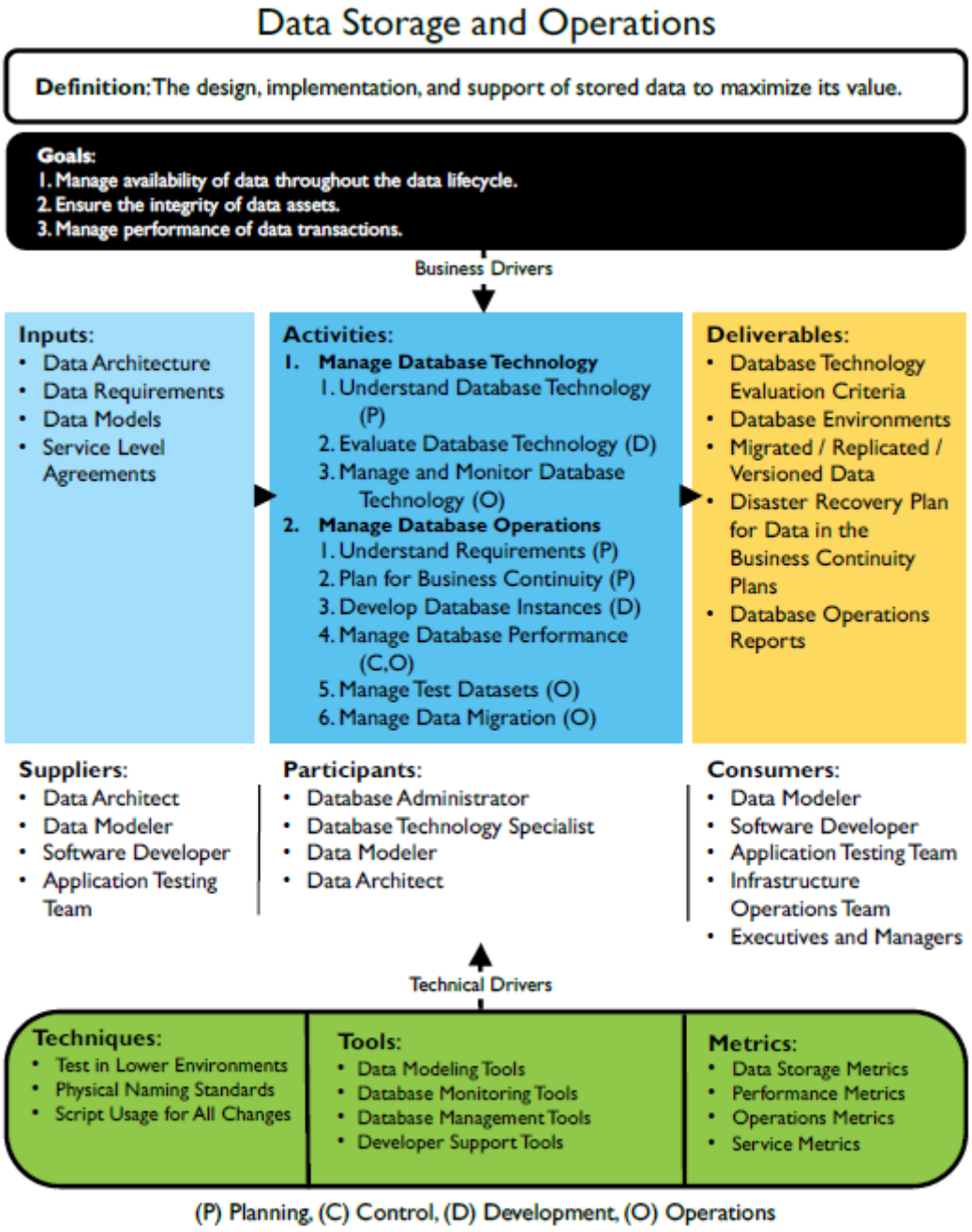

Data Storage and Operations
Definition: The design, implementation, and support of stored data to maximize its value.
Goals:
1. Manage availability of data throughout the data lifecycle.
2. Ensure the integrity of data assets.
3. Manage performance of data transactions.
Business Drivers
Inputs:
• Data Architecture
• Data Requirements
• Data Models
• Service Level Agreements
Activities:
1. Manage Database Technology
1. Understand Database Technology (P)
2. Evaluate Database Technology (D)
3. Manage and Monitor Database Technology (O)
2. Manage Database Operations
1. Understand Requirements (P)
2. Plan for Business Continuity (P)
3. Develop Database Instances (D)
4. Manage Database Performance (C,O)
5. Manage Test Datasets (O)
6. Manage Data Migration (O)
Deliverables:
• Database Technology Evaluation Criteria
• Database Environments
• Migrated / Replicated / Versioned Data
• Disaster Recovery Plan for Data in the Business Continuity Plans
• Database Operations Reports
Suppliers:
• Data Architect
• Data Modeler
• Software Developer
• Application Testing Team
Participants:
• Database Administrator
• Database Technology Specialist
• Data Modeler
• Data Architect
Consumers:
• Data Modeler
• Software Developer
• Application Testing Team
• Infrastructure Operations Team
• Executives and Managers
Technical Drivers
Techniques:
• Test in Lower Environments
• Physical Naming Standards
• Script Usage for All Changes
Tools:
• Data Modeling Tools
• Database Monitoring Tools
• Database Management Tools
• Developer Support Tools
Metrics:
• Data Storage Metrics
• Performance Metrics
• Operations Metrics
• Service Metrics
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 54 Context Diagram: Data Storage and Operations

Data Storage and Operations includes the design, implementation and support of stored data, to maximise its value throughout the lifecycle, from creation to disposal. Two sub-activities:

- **Database support:** Activities related to the data lifecycle from implementation of a database environment, through obtaining, backing up and purging data. It includes ensuring the database performs well. Monitoring and tuning
- **Database technology support:** Defining technical requirements that meet organisational needs, defining technical architecture, installing, maintaining technology, and resolving related issues.

1.1 Business Drivers

Business continuity

1.2 Goals and principles

Goals:
1. Manage availability of data throughout the data lifecycle.
2. Ensure the integrity of data assets.
3. Manage performance of data transactions.

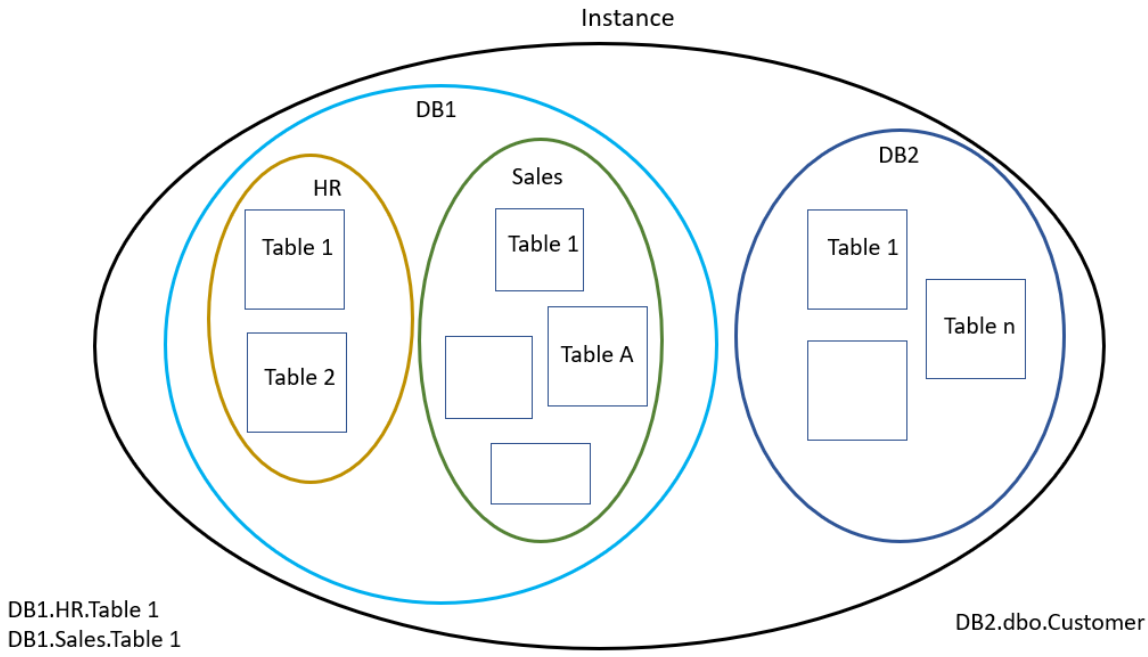
Highly technical side of data management. Guiding principles:

- **Identify and act on automation opportunities:** Automate database processes, develop tools and processes that shorten cycles, reduce errors and development rework. Do in collaboration with data modelling and Data Architecture.
- **Build with reuse in mind:** Develop abstracted and reusable data objects
- **Understand and appropriately apply best practices:**
- **Connect database standards to support requirements:** SLAs reflect DBA-recommended and developer-accepted methods of ensuring integrity and data security.
- **Set expectation for the DBA role in project work:** Include the DBA in all SDLC phases of a project

1.3 Essential Concepts

1.3.1 Database Terms

- **Database:** Any collection of stored data
- **Instance:** An execution of database software controlling access to a certain area of storage
- **Schema:** A subset of database objects contained within a database or instance, usually with something in common. Used to isolate sensitive data from general user base
- **Node:** An individual computer part of a distributes database
- **Database abstraction:** An API (common Application Interface) is used to call database functions



Explanation of the Relationship Between Database Instance, Database, Schema and Database Objects

1.3.2 Data Lifecycle Management

The DBA is the custodian of all database changes. The DBA defines the precise changes, implements and controls the changes. The DBA must have a plan to back out the changes.

1.3.3 Administrators

- **Production DBA:** Responsible for data operations management
  - Ensures performance and reliability of the database – performance tuning, monitoring, error reporting
  - Backup and recovery mechanisms
  - Implement clustering and failover of the database if continual data availability is required
  - Other database maintenance activities such as archiving
  - Deliverables of Production DBAs:
    - A production database environment
    - Mechanisms and processes for controlled implementation of changes to databases in the production environment
    - Mechanisms for ensuring availability, integrity and recoverability of data in response to all circumstances that could result in loss or corruption of data.
    - Mechanisms for detecting and reporting database errors
    - Data availability, recovery and performance in accordance with SLAs
    - Mechanisms and processes for monitoring database performance
- **Application DBA:** Responsible for databases in all environments (Development/test, QA and production). Part of an application support team
- **Procedural and Development DBAs:**
  - Administration of procedural database objects (stored procedures, triggers and user-defined functions)
  - Development DBAs focus on creating and managing special use databases such as "sandbox"

- **NSA:** Network Storage Administrators support data storage arrays

1.3.4 Database Architecture Types

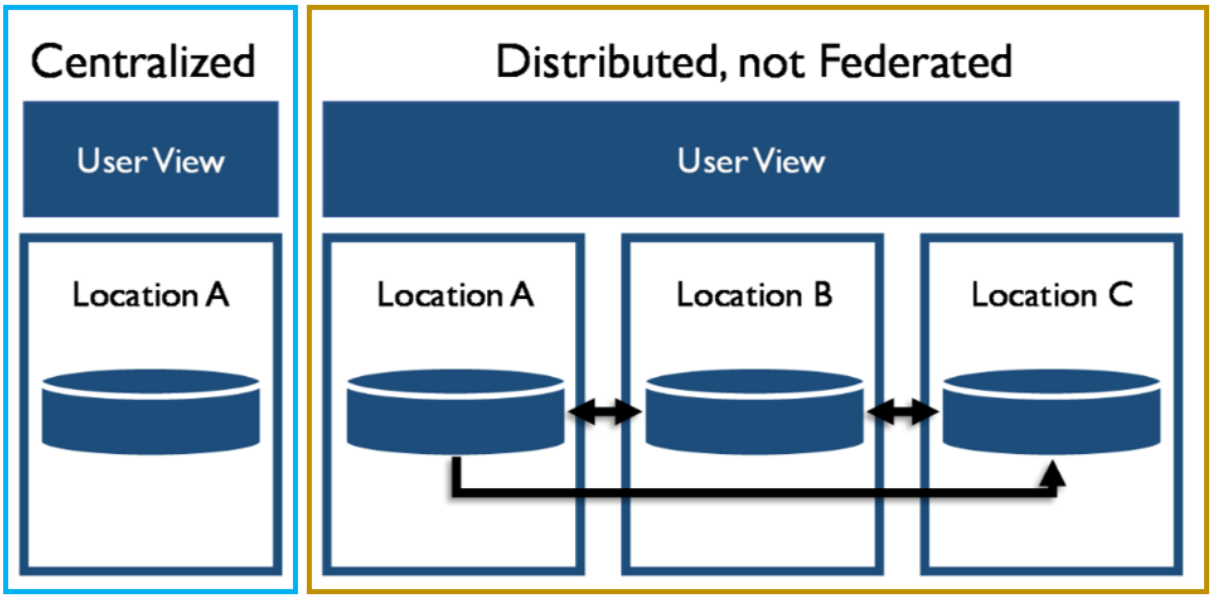


Figure 55 Centralised vs Distributed

1.3.4.1 Centralised Databases

All the data in one place. All users come to the one system, no alternatives if it is unavailable.

1.3.4.2 Distributed Databases

Quick access to data over many of nodes, each offering local computation and storage. DBMS replicates across nodes, and can detect and handle failures. MapReduce divides a data request into small fragments over the nodes.

1.3.4.2.1 Federated Databases

A federated database system maps multiple autonomous database systems into a single federated database, connected via a network. Remain autonomous but allow sharing of data.

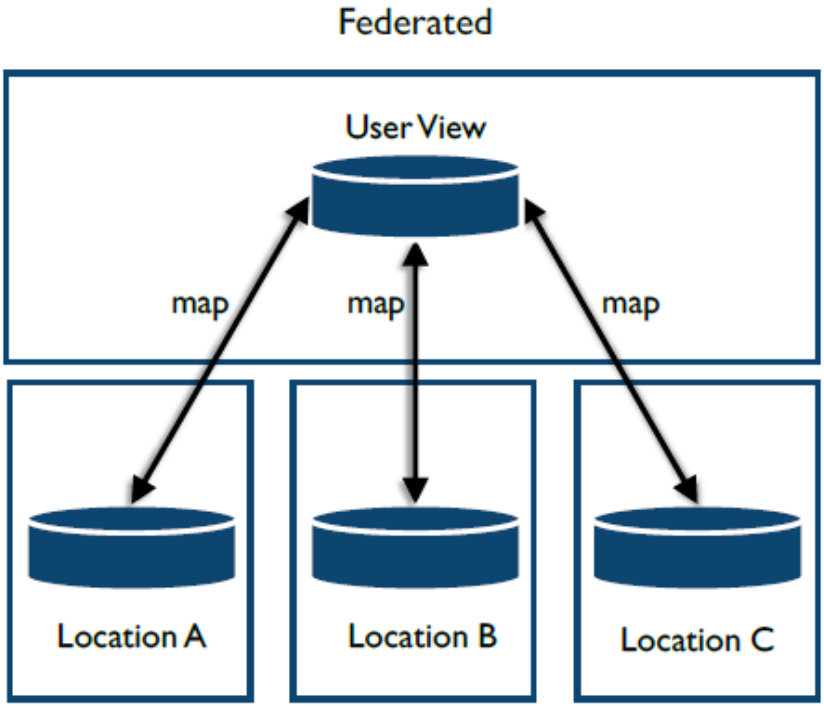


Figure 56 Federated Databases

A FDBMS can be loosely or tightly coupled:

- **Loosely coupled**: Component databases construct their own federated schema. Multi-database language and user knows the schema
- **Tightly coupled**: Component systems use independent processes to construct and publish an integrated federated schema (Note: the diagram labels on Figure 57 Coupling in the DMBOK V2 are switched)

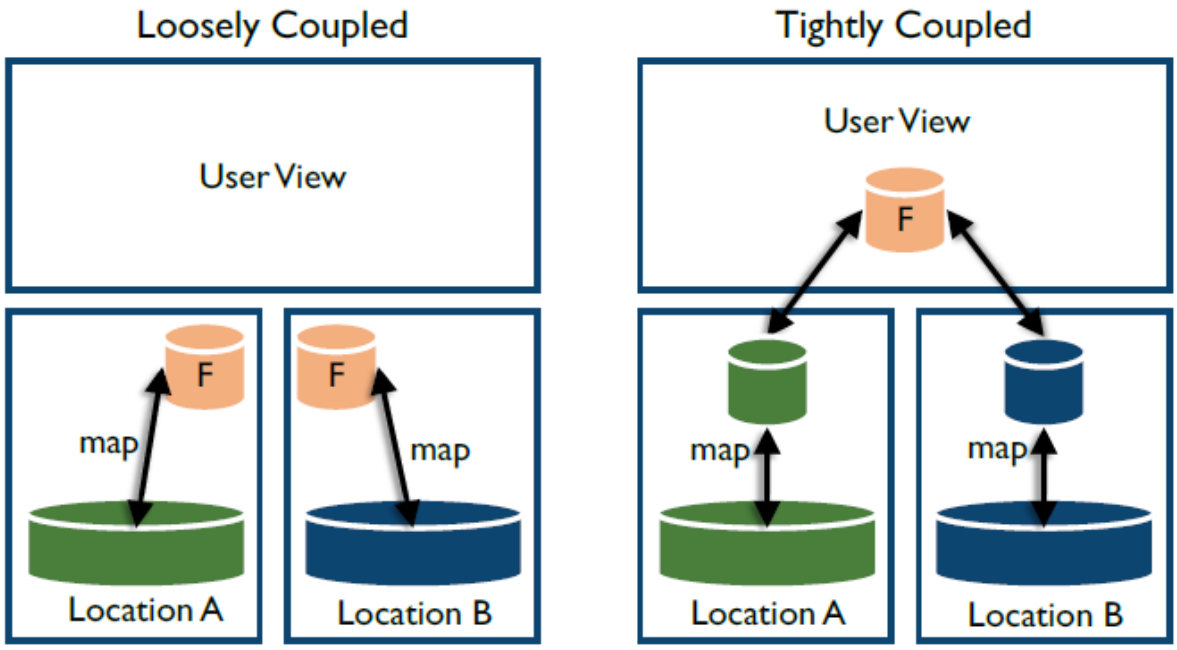
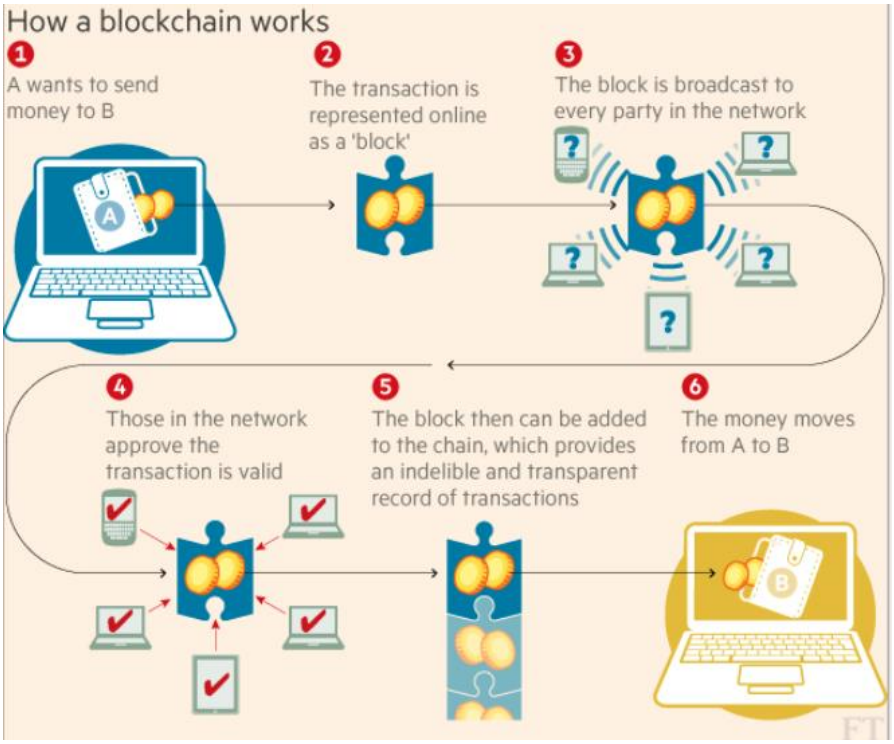


Figure 57 Coupling

1.3.4.2.2 Blockchain database

A type of federated database used to securely manage financial transactions. Each transaction has a record. The database creates chains of time bound groups (blocks) that also contain information from the previous block in the chain. Hash algorithms create information on the transactions which can never change. Tampering is evident if the hash values no longer match.


How a blockchain works
1
A wants to send money to B
2
The transaction is represented online as a 'block'
3
The block is broadcast to every party in the network
4
Those in the network approve the transaction is valid
5
The block then can be added to the chain, which provides an indelible and transparent record of transactions
6
The money moves from A to B

Blockchain Database Explained (Author Unknown)

1.3.4.3 Virtualisation / Cloud platforms

Virtualisation (Cloud computing) provides computation, software, data access and storage without end-user knowledge of the physical location or configuration of the system.

- **Virtual machine image:** Users purchase virtual machine instances for a limited time. Either load own database or use an optimised installation of the database
- **Database-as-a-service (DaaS):** Database service provider installs and maintains the database and application owners pay according to their usage
- **Managed database hosting in the cloud:** The cloud provider hosts the database and manages it on the application owner's behalf

DBAs and Network and System Administrators need to establish a systematic project approach to the following functions (as well as the security aspects):

- **Standardisation/consolidation:** Based on Data Governance policy, consolidation reduces the number of data storage locations and standard procedures are developed by DBAs and Data Architects
- **Server virtualisation:** Enables reduction of cost in infrastructure management
- **Automation** of data tasks
- **Security:** Integrate security of virtual systems with security of physical infrastructures

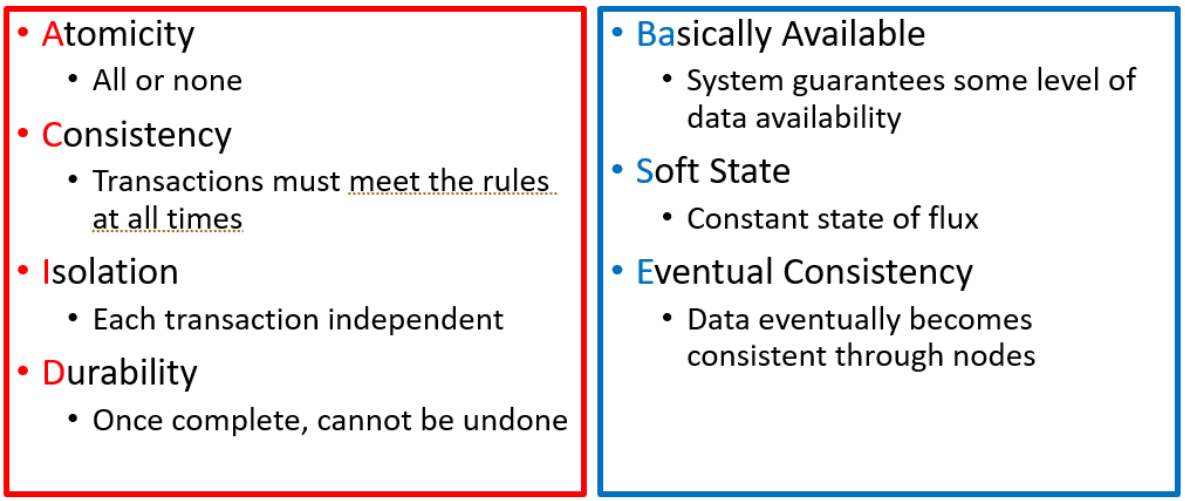
1.3.5 Database Processing Types

ACID: 1980s – reliability within database transactions. Relational – SQL Server.

- **Atomicity:** All operations are performed, or none are. If one fails, the entire transaction fails
- **Consistency:** The transaction must meet all rules defined by the system at all times and must void a half-completed transaction
- **Isolation:** Each transaction is independent
- **Durability:** Once complete, the transaction cannot be undone

BASE: Increase in volumes and variability of data, the need to document and store **l**ess structured data, read-optimised data workloads, greater flexibility of scaling and design

- **Basically available:** System guarantees some level of data availability even if there are node failures
- **Soft state:** System in constant state of flux. Data may be available but not current
- **Eventual consistency:** Data eventually becomes consistent through nodes



Summary of ACID and BASE Acronyms (Slide in Data Management for Certification Course)

Item	ACID	BASE
Casting (data structure)	Schema must exist	Dynamic
	Table structure exists	Adjust on the fly
	Columns data typed	Store dissimilar data
Consistency	Strong Consistency Available	Strong, Eventual, or None
Processing Focus	Transactional	Key-value stores
Processing Focus	Row/Column	Wide-column stores
History	1970s application storage	2000s unstructured storage
Scaling	Product Dependent	Automatically spreads data across commodity servers
Origin	Mixture	Open-source
Transaction	Yes	Possible

Table 12 ACID vs BASE

- **CAP:** CAP Theorem (or Brewers Theorem) – the larger the distributed system, the lower the compliance to ACID. There is a trade-off between the following properties and at most 2 of the following 3 properties can exist in a shared-data system:
    - **Consistency:** The system must operate as designed and expected at all times
    - **Availability:** The system must be available and respond to each request

- **Partition Tolerance:** The system must be able to continue operations during occasions of data loss or partial system failure

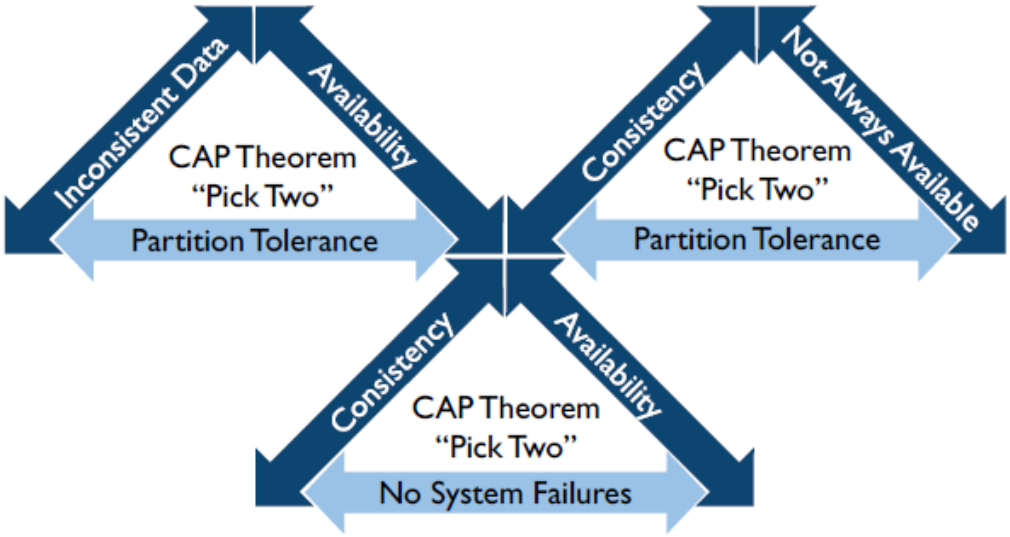


Figure 58 CAP Theorem

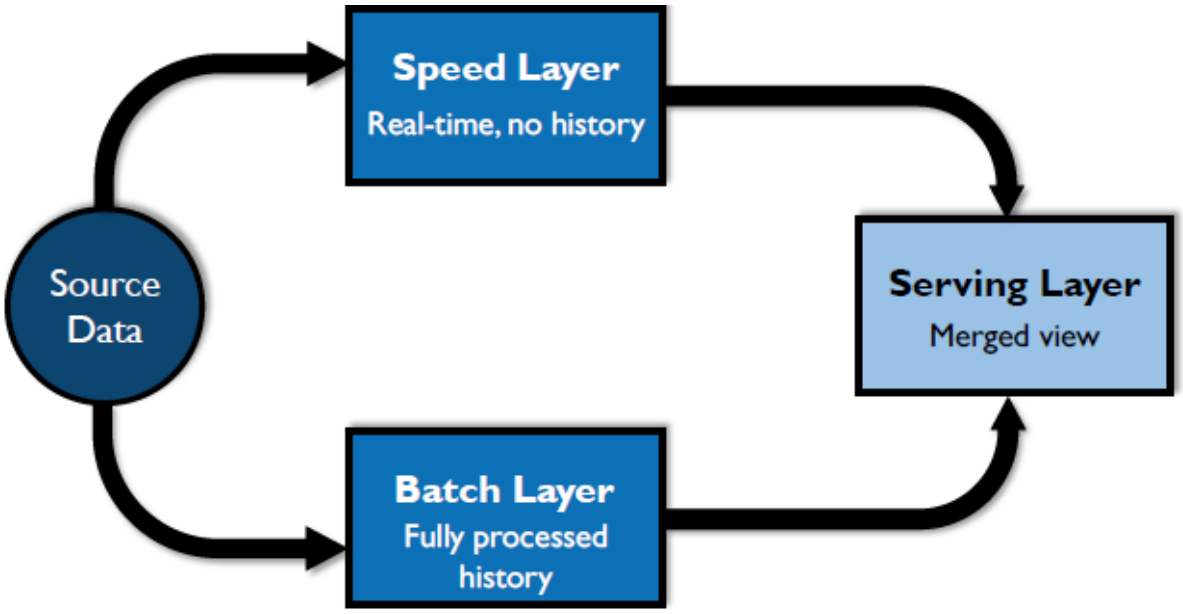


Figure 101 Services-based Architecture

Used to drive Lambda Architecture for big data, which uses two paths for data: (refers to Chapter 14 where it appears as 14.1.3.7 Services-Based Architecture (SBA)):

- Speed path: Availability and Partition Tolerance are most important
- Batch path: Consistency and Availability are most important

1.3.6 Data Storage Media

- **Disk and Storage Area Networks (SAN):** Disk arrays and SAN. Persistent storage
- **In-Memory:** In-Memory Databases (IMDB) are loaded into volatile memory where all processing takes place. Faster response than disk

- **Columnar Compression Solution:** For data sets where data values are repeated to a large extent and may be compressed. Reduces I/O time
- **Flash Memory:** Solid state

1.3.7 Database Environments

There are various environments used during the system development lifecycle.

- **Production Environments:** The technical environment where all business processes occur. The DBA team should be the only team implementing changes, adhering strictly to standards and procedures
- **Pre-Production Environments:**
  - **Development:** Slimmer version of production for developers to test new code or patches
  - **Test:** Used for:
    - Quality Assurance testing (QA)
    - Integration testing
    - User Acceptance Testing (UAT)
    - Performance Testing
  - **Sandboxes or Experimental Environments:** Can be read only access to production data for experimentation

1.3.8 Database Organisation

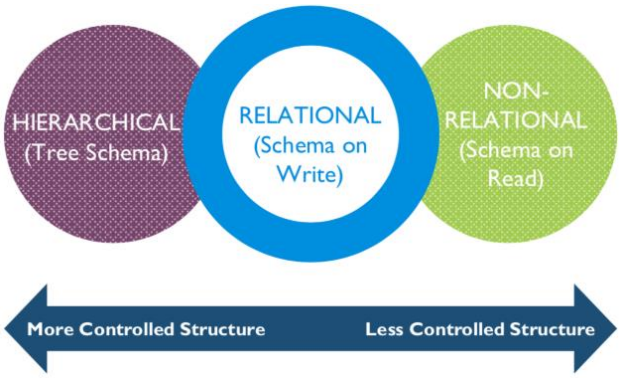
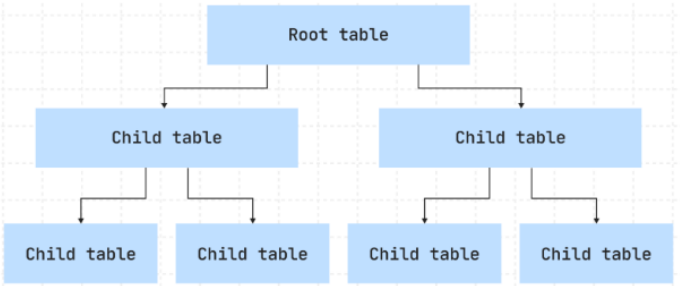


Figure 59 Database Organisation Spectrum

1.3.8.1 Hierarchical

Oldest database model. Tree structure with one-to-many parent-child relationships. Example Directory trees and XML.



1.3.8.2 Relational

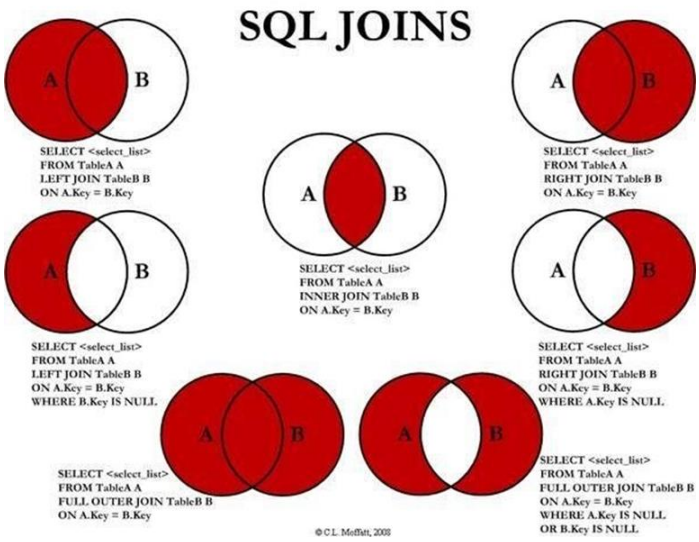
Based on set theory and relational algebra. Set operations (union, intersection, minus) in the form of SQL (Structured Query Language) are used to retrieve data. To write data the schema must be known (schema on write). Relational databases are row-oriented. The database management system is called RDBMS

- Tables are sets or relations
- Attributes (columns)
- Tuples (Rows)

SQL uses set operations to retrieve data

Variations on relational databases:

- **Multidimensional:** Allows searching using several data element filters simultaneously. Uses a type of SQL called MDX (Multidimensional eXpression)
- **Temporal:** Built in support for handling time



1.3.8.3 Non-relational (NoSQL – Not only SQL)

Schema-on-read allows data to be read in different ways. NoSQL means the storage structure is not bound by tabular design. NoSQL databases are used in Big Data and real-time web applications as they are optimised for retrieval and appending operations.

- **Column-oriented:** Used in BI applications as redundant data can be compressed. Difference between column-oriented and row-oriented (usually relational):
  - **Column-oriented is more efficient when:**
    - Aggregating over many rows
    - New values for that column are applied at once as there is no need to touch columns for the other rows
    - Online Analytical Processing (OLAP)
  - **Row-oriented is more efficient when:**
    - Many columns of a single row required at once
    - Writing a whole row of new data
    - Online Transaction Processing (OLTP)
- **Spatial:** Store and query data that represents objects defined in geometrical space. Use special indexes to perform database operations. Open Geospatial Consortium standard
- **Object/Multi-media:** Hierarchical storage Management System manages the media objects
- **Flat File Database:** Data in rows and columns as a single file. Used by Hadoop databases
- **Key-Value Pair:** A key identifier and a value:
  - **Document Databases:** Each document has a key. Use XML or JSON (Java Script Object Notation) structures
  - **Graph Databases:** Key-value pairs where the focus is on the relationship between nodes rather than the nodes themselves
- **Triplestore:** A data entity composed of subject-predicate-object. Best for taxonomy and thesaurus management. Resource Description Framework (RDF):
  - **Subject:** a resource
  - **Predicate:** relationship between the subject and object

- **Object:** the object

1.3.9 Specialised Databases

- Computer Assisted Design and Manufacturing (CAD/CAM): Object database
- Geographical Information Systems (GIS): Geospatial databases
- Shopping-cart applications: XML databases to store customer order data

1.3.10 Common Database Processes

1.3.10.1 Archiving

The process of moving data off immediately accessible media onto less expensive media with lower retrieval performance. Schedule regular restoration tests.

Archival processes must be aligned with the partitioning strategy to ensure optimal availability and retention:

- Create a secondary area on a secondary database server
- Partition database tables into archival blocks
- Replicate to the separate database
- Create backups (tape or disk)
- Create jobs that periodically purge unneeded data

1.3.10.2 Capacity and growth projections

Decide how much storage is needed, and work out expansion needs.

1.3.10.3 Change Data Capture (CDC)

Detecting that data has changed and ensure information relevant to the change is stored appropriately. Log-based replication.

1.3.10.4 Purging

Purging is the process of completely removing data from media so that it cannot be retrieved.

1.3.10.5 Replication

Replication transparency: The same data is stored on multiple storage devices and is consistent throughout the database system, so that users cannot tell which database copy they are using.

- **Active replication:** Create and store the same data at every replica from every other replica
- **Passive replication:** recreating and storing data on the primary, then transferring it to secondary replicas

Replication has 2 dimensions of scaling:

- Horizontal: More data replicas
- Vertical: Data replicas located geographically further away

Multi-master replication:

- Update any node and ripples through other servers
- Often desired, but increases complexity and cost

Two primary replication patterns:

- **Mirroring:** Updates to the primary are replicated immediately to the secondary as part of a two-phase commit**.**

- **Log Shipping:** The secondary receives and applies copies of the primary database's transaction log at regular intervals

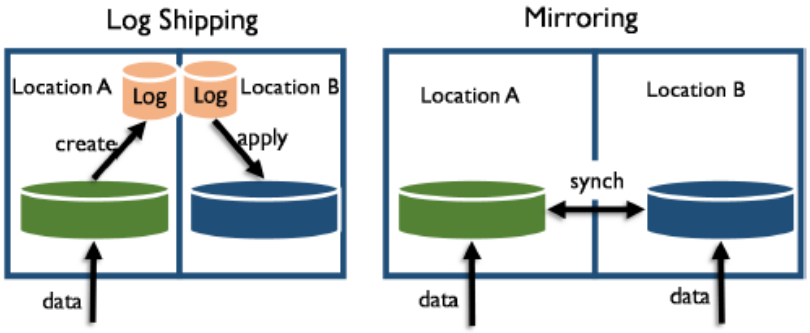


Figure 60 Log Shipping vs Mirroring

1.3.10.6 Resiliency and Recovery

Resiliency is a measure of how tolerant a system is to error conditions, and how well it recovers. Three recovery types:

- **Immediate recovery:** Predicting and automatically resolving issues
- **Critical recovery:** Plan to restore the system quickly to minimise delays to business processes
- **Non-critical recovery**: Restoration can be delayed until critical systems have been restored

1.3.10.7 Retention

How long data is kept available. Affects capacity planning. Data retention plans are also affected by Data Security, as some data has legal requirements to be retained a specific time.

1.3.10.8 Sharding

Small chunks of the database are isolated so replication is merely a file copy.

2 Activities

Data Technology Support (selecting and maintaining the software that stores and manages the data) and Data Operations Support (the data and processes that the software manages).

2.1 Manage Database Technology

The **Information Technology Infrastructure Library (ITIL)** is the leading reference model.

2.1.1 Understand Database Technology

DBAs and Data Architects combine their knowledge of available tools with business requirements to suggest the best technology.

2.1.2 Evaluate Database Technology

Some of the factors to consider when selecting a DBMS:

- Product architecture and complexity
- Volume and velocity limits
- Application profile such as transaction processing and BI
- Hardware platform and operating system support
- Availability of supporting software tools
- Performance benchmarks

- Scalability
- Software, memory and storage requirements
- Resiliency, including error handling and reporting

Price and the possible necessity to employ extra staff.

2.1.3 Manage and Monitor Database Technology

DBAs need to be trained to function as Level 2 technical support. They need to have a working knowledge of application development skills. DBAs work with business users and application developers to ensure the most effective use of technology.

2.2 Manage Database Operations

Database support is provided by DBAs and NSAs (Network Storage Administrators).

2.2.1 Understand Requirements

- **Define storage requirements:** Initial capacity estimate for first year and growth projection for next few years. Take data storage, indexes, logs and mirrors into account
- **Identify usage patterns:** Databases have predictable usage patterns:
  - Transaction based
  - Large data set write or retrieval
  - Time based – month-end, lighter on weekends
  - Location based – more populated areas have more transactions
  - Priority based – some departments or batch IDs have higher priority
- **Define access requirements:** The authorisation to access different data files, and the standard languages and methods to do it

2.2.2 Plan for Business Continuity

Recovery plan for all databases and database servers in the event of a disaster which could result in loss or corruption of data. Identify critical databases which need to be restored first.

This plan should be reviewed by the business continuity group.

- **Make backups:**
  - Back up databases and transaction logs
  - Backup frequency determined by SLA.
  - Incremental and complete backups
  - Keep backups on a separate file system
- **Recover data:** DBA executes restoration of data. Test recovery periodically

2.2.3 Develop Database Instances

- Installing and updating DBMS software
- Maintaining multiple environment installations, including different DBMS versions
- Installing and administering related data technology

2.2.3.1 Manage the Physical Storage Environment

Storage environment management needs to follow Software Configuration Management (SCM) processes or ITIL methods to record modification to the database configuration. Four processes:

- **Configuration Identification:** DBAs with Data Stewards, Data Architects and Data Modellers to identify attributes that define end-user configuration. These must be baselined, recorded and only changed with formal change control

- **Configuration change control:** Processes and approval stages to change the above attributes
- **Configuration status accounting:** Report on the configuration at any point in time
- **Configuration audits:**
    - Physical configuration audit: an item is installed in accordance with design documentation
    - Functional configuration audit: Performance attributes of an item are achieved

DBAs must communicate any changes to the physical attributes to modellers, developers and Metadata managers.

DBAs also maintain metrics on data volume, capacity projections, query performance and statistics on the physical objects.

2.2.3.2 Manage Database Access Controls

DBAs oversee the following functions to protect data assets:

- **Controlled Environment:** DBAs and NSAs. Network roles and permissions, 24/7 monitoring, firewall management, patch management
- **Physical security:** Simple Network Management Protocol (SNMP)-based monitoring, data audit logging, disaster management and database backup plans
- **Monitoring:** Continuous monitoring of servers
- **Controls:** Access controls, database auditing, intrusion detection and vulnerability assessment tools

2.2.3.3 Create Storage Containers

All data must be stored on the physical drive and organised for ease of load, search and retrieval.

2.2.3.4 Implement Physical Data Models

DBAs implement the physical layout of the data model in storage. The physical data model includes storage objects, indexing objects, and any encapsulated code objects required to enforce quality rules, connect database objects and achieve database performance.

2.2.3.5 Load Data

DBMS should have a bulk load facility to load data into a new database. Data must be in the right format for the target object.

Third party data can be updated regularly by the service. DBAs must be aware of any legal restrictions before loading third party data.

DBAs may load data manually, or automate and schedule loading.

Responsibility for data acquisition services in a managed environment is centralised with data analysts who document it in the logical data model. Developers create scripts if necessary. DBA implements the processes to load the data into the database.

2.2.3.6 Manage Data Replication

DBAs advise data replication decisions on:

- Active or passive replicating
- Distributed concurrency control from distributed data systems
- The appropriate methods to identify updates to data under the Change Data Control process:
    - Timestamp

- Version numbers

2.2.4 Manage Database Performance

Database performance depends on availability and speed. Availability of space and query optimisation are part of performance.

2.2.4.1 Set Database Performance Service Levels

System performance, data availability and recovery expectations, and expectations for teams to respond are governed by Service Level Agreements (SLAs) between IT data management services and data owners.

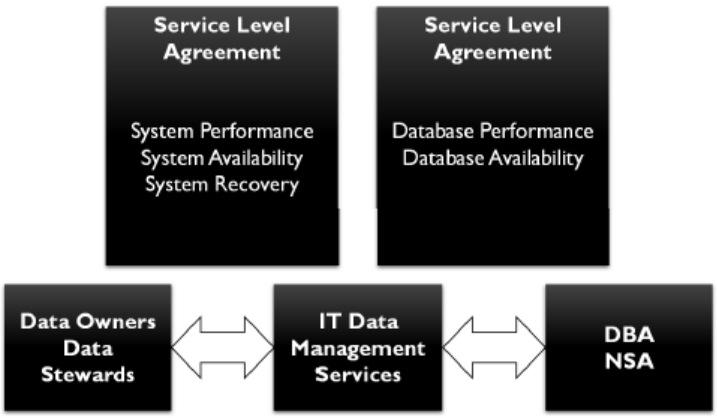


Figure 61 SLAs for System and Database Performance

2.2.4.2 Manage Database Availability

Availability is the percentage of time that a system or database is available for productive work. Four factors affect availability:

- **Manageability:** The ability to create and maintain an environment
- **Recoverability:** Establish service after interruption, and correct the errors caused
- **Reliability:** Ability to deliver service at specified levels for a stated period
- **Serviceability:** Identify and diagnose problems and solve them

Some things that prevent databases from being available:

- Planned outages
- Unplanned outages
- Application problems
- Data problems
- Human error

DBAs are responsible for doing everything possible to ensure databases stay online and operational:

- Run database backup utilities
- Run database reorganisation utilities
- Run statistics gathering utilities
- Run integrity gathering utilities
- Automating the above utilities
- Exploiting table clustering and partitioning
- Replicating data across mirror databases to ensure high availability

2.2.4.3 Manage Database Execution

DBAs manage database execution, logging and log sizes and synchronisation.

2.2.4.4 Maintain database Performance Service Levels

DBAs generate performance analysis reports regularly and compare them with previous reports to identify negative trends and analyse problems over time.

- **Transaction performance vs Batch performance**: Batch jobs must complete within a batch window
- **Issue remediation:** Common reasons for poor database performance are:
  - Memory allocation or contention
  - Locking and blocking
  - Inaccurate database statistics
  - Poor coding
  - Inefficient complex table joins
  - Insufficient indexing
  - Application activity
  - Overloaded servers
  - Database volatility
  - Runaway queries

2.2.4.5 Maintain Alternate Environments

Types of alternate environments:

- **Development:** Test changes that will be implemented in production
- **Test:** QA, Integration testing, UAT and performance testing
- **Sandboxes:** Experimental environments
- **Alternate production environments:** Support failover, offline backups and resiliency support systems

2.2.5 Manage Test Data Sets

Efficient testing requires high quality test data to be generated and managed.

2.2.6 Manage Data Migration

Data Migration is the process of transferring data between storage types, formats or computer systems with as little change as possible. Usually performed programmatically, automated based on rules.

3 Tools

- Data modelling tools
- Database monitoring tools
- Database Management tools
- Developer support tools

4 Techniques

- **Test in lower environments:**
  - Test upgrades and patches on the lowest level first, development
  - Then install and test on higher levels, production last
- **Physical Naming Standards:** ISO/IEC 11179 – Metadata Registries (MDR) Part 5

- **Script Usage for All Changes:** Test any change scripts in non-production before applying

5 Implementation Guidelines

5.1 Readiness assessment/Risk assessment

Two central ideas:

- **Data Loss**: SLAs specify general requirements for protection. Ongoing assessment to ensure robust technical responses to cyber threat are in place
- **Technology readiness**: Does the organisation have the skill set to implement newer technology?

5.2 Organisation and Cultural Change

- **Proactively communicate**: DBAs should be in close communication with project teams at all stages of the project, to detect and resolve issues as early as possible
- **Communicate with people on their level and in their terms**
- **Stay business focussed:** Objective is to meet business requirements and derive maximum value
- **Be helpful:** Not helping may force people to ignore standards and find another way
- **Learn continually:** Setbacks and problems are lessons which can be applied later

Understand stakeholders and their needs. Develop clear, concise, practical, business focussed standards for doing the best possible work in the best possible way. Teach and implement those standards in a way that provides maximum value to stakeholders and earns their respect.

6 Data Storage and Operations Governance

6.1 Metrics

Storage metrics may include:

- Count of databases by type
- Aggregated transaction stats
- Capacity metrics
- Storage service usage
- requests made against storage services
- Performance improvements of applications using service

Performance metrics:

- Transaction frequency and quantity
- Query performance
- API service performance

Operational metrics:

- Aggregated statistics about data's retrieval times
- Backup size
- Data quality measurement
- Availability

Service metrics:

- Issue submission, resolution and escalation count by type
- Issue resolution time

6.2 Information asset Tracking

- Ensure organisation complies with **software licensing and annual support agreements**
- Determine **TCO (total cost to ownership)** for each technology product

6.3 Data Audits and Data validation

A **data audit** is the evaluation of data based on defined criteria, typically performed to investigate specific concerns about the data. A data audit includes:

- Project specific checklist
- Comprehensive checklist
- Required deliverables
- Quality control criteria

Data validation is the process of evaluating stored data against established acceptance criteria (from the Data Quality team or customer specifications) to determine its quality and usability.

DBAs support data audits and validation by:

- Help develop and review the approach
- Perform preliminary data screening and review
- Develop data monitoring methods
- Applying statistical, geo-statistical and bio-statistical techniques to optimise the data
- Support sampling and analysis
- Review data
- Provide support for data discovery
- Act as the SME for questions related to database administration

Chapter 7
Data Security

1 Introduction

Effective data security policies and procedures ensure the right people can use and update data the right way, and all inappropriate access and update is restricted (Ray, 2012).

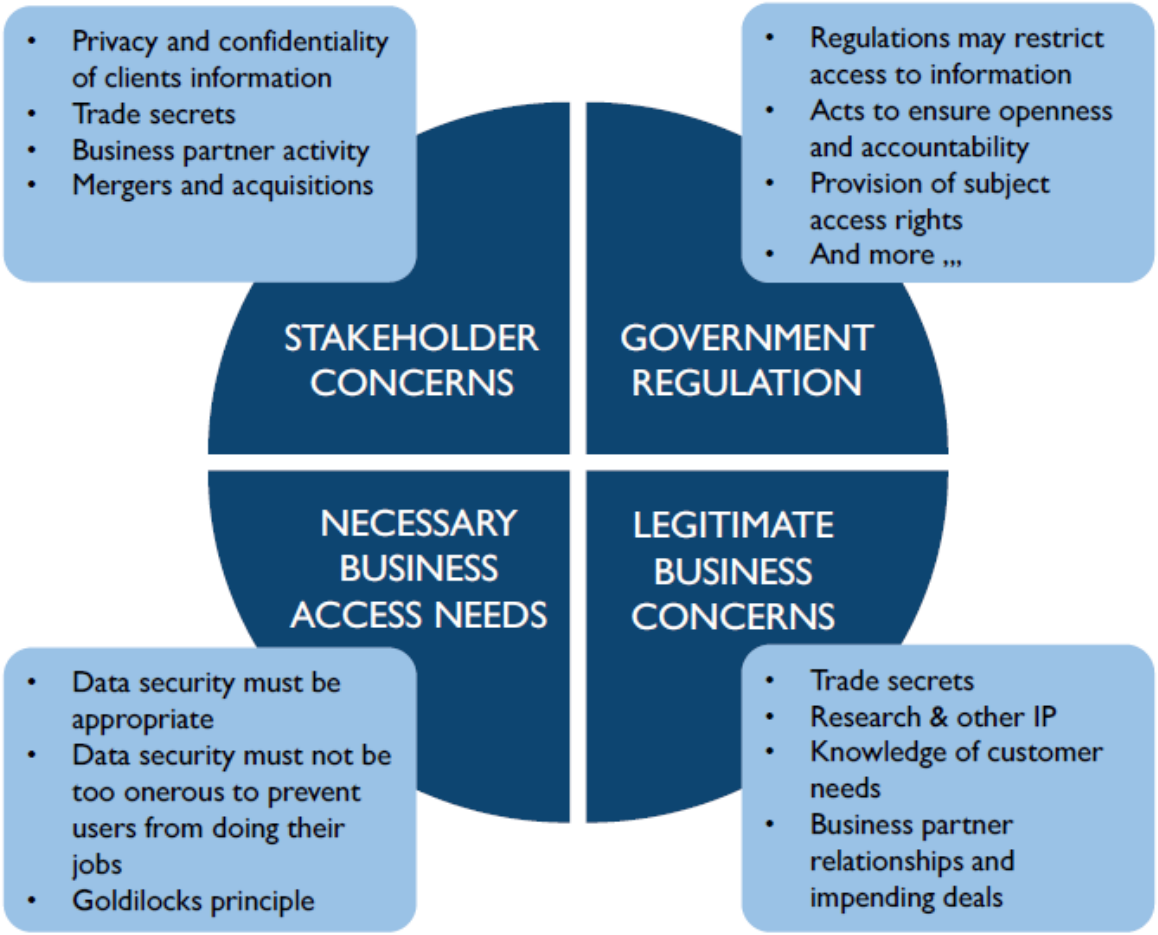

Privacy and confidentiality of clients information
Trade secrets
Business partner activity
Mergers and acquisitions
Regulations may restrict access to information
Acts to ensure openness and accountability
Provision of subject access rights
And more ,,,
STAKEHOLDER CONCERNS
GOVERNMENT REGULATION
NECESSARY BUSINESS ACCESS NEEDS
LEGITIMATE BUSINESS CONCERNS
Data security must be appropriate
Data security must not be too onerous to prevent users from doing their jobs
Goldilocks principle
Trade secrets
Research & other IP
Knowledge of customer needs
Business partner relationships and impending deals

Figure 62 Sources of Data Security Requirements

- **Stakeholders:** Everyone in an organisation must be a responsible trustee of stakeholders' data
- **Government regulations:** Some restrict access and others ensure openness, transparency and accountability
- **Proprietary business concerns:** Data which provides competitive advantage must be protected
- **Legitimate access needs**
- **Contractual obligations:** Contractual and non-disclosure agreements

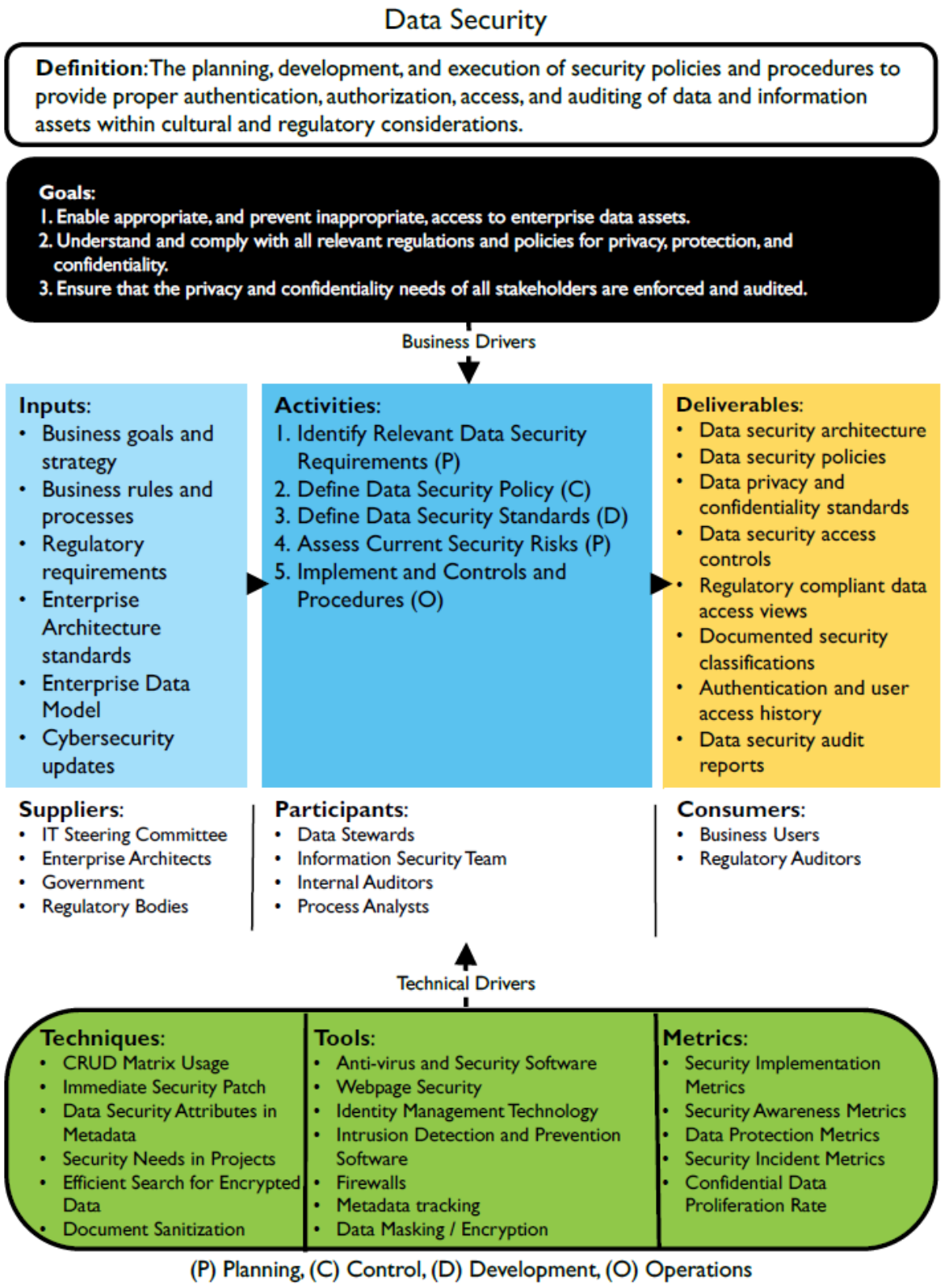

Data Security
Definition: The planning, development, and execution of security policies and procedures to provide proper authentication, authorization, access, and auditing of data and information assets within cultural and regulatory considerations.
Goals:
1. Enable appropriate, and prevent inappropriate, access to enterprise data assets.
2. Understand and comply with all relevant regulations and policies for privacy, protection, and confidentiality.
3. Ensure that the privacy and confidentiality needs of all stakeholders are enforced and audited.
Business Drivers
Inputs:
• Business goals and strategy
• Business rules and processes
• Regulatory requirements
• Enterprise Architecture standards
• Enterprise Data Model
• Cybersecurity updates
Activities:
1. Identify Relevant Data Security Requirements (P)
2. Define Data Security Policy (C)
3. Define Data Security Standards (D)
4. Assess Current Security Risks (P)
5. Implement and Controls and Procedures (O)
Deliverables:
• Data security architecture
• Data security policies
• Data privacy and confidentiality standards
• Data security access controls
• Regulatory compliant data access views
• Documented security classifications
• Authentication and user access history
• Data security audit reports
Suppliers:
• IT Steering Committee
• Enterprise Architects
• Government
• Regulatory Bodies
Participants:
• Data Stewards
• Information Security Team
• Internal Auditors
• Process Analysts
Consumers:
• Business Users
• Regulatory Auditors
Technical Drivers
Techniques:
• CRUD Matrix Usage
• Immediate Security Patch
• Data Security Attributes in Metadata
• Security Needs in Projects
• Efficient Search for Encrypted Data
• Document Sanitization
Tools:
• Anti-virus and Security Software
• Webpage Security
• Identity Management Technology
• Intrusion Detection and Prevention Software
• Firewalls
• Metadata tracking
• Data Masking / Encryption
Metrics:
• Security Implementation Metrics
• Security Awareness Metrics
• Data Protection Metrics
• Security Incident Metrics
• Confidential Data Proliferation Rate
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 63 Context Diagram: Data Security

1.1 Business Drivers

Data security risks are associated with regulatory compliance, fiduciary responsibility, reputation and a legal and moral responsibility to protect private and sensitive data. Mitigating risks and growing the business can be complimentary when integrated into an information protection strategy.

- **Risk reduction:** Should be enterprise wide. Classify organisation's data:
  - Identify and classify sensitive data assets
  - Locate sensitive data throughout the enterprise
  - Determine how each asset needs to be protected
  - Identify how the information interacts with business processes
  - Identify external (hackers) and internal (employees and processes) threats
- **Business Growth:** Data security breaches can impact business growth. Robust data security inspires consumer confidence. Trusted e-commerce drives profit and growth
- **Security as an asset:** Tag data with security Metadata

1.2 Goals and Principles

Goals:
1. Enable appropriate, and prevent inappropriate, access to enterprise data assets.
2. Understand and comply with all relevant regulations and policies for privacy, protection, and confidentiality.
3. Ensure that the privacy and confidentiality needs of all stakeholders are enforced and audited.

Principles:

- **Collaboration**: Data Security is a collaborative effort involving IT security administrators, data stewards/data governance, internal and external audit teams, and the legal department
- **Enterprise approach**: Data Security standards and policies must be applied consistently across the entire organization
- **Proactive management**: Success in data security management depends on being proactive and dynamic, engaging all stakeholders, managing change, and overcoming organizational or cultural bottlenecks such as traditional separation of responsibilities between information security, information technology, data administration, and business stakeholders
- **Clear accountability**: Roles and responsibilities must be clearly defined, including the 'chain of custody' for data across organizations and roles
- **Metadata-driven**: Security classification for data elements is an essential part of data definitions
- **Reduce risk by reducing exposure**: Minimize sensitive/confidential data proliferation, especially to non-production environments

1.3 Essential Concepts

1.3.1 Vulnerability

A **vulnerability** is a weakness or defect in a system that allows it to be successfully attacked and compromised – essentially a hole in an organization's defences. Some vulnerabilities are called **exploits.**

1.3.2 Threat

A **threat** is a potential offensive action that could be taken against an organization. Threats can be internal or external. They are not always malicious. A uninformed insider can take offensive actions again the organization without even knowing it.

1.3.3 Risk

The term **risk** refers both to the possibility of loss and to the thing or condition that poses the potential loss. Risk calculations:

- Probability that the threat will occur and its likely frequency
- The type and amount of damage created each occurrence might cause, including damage to reputation
- The effect damage will have on revenue or business operations
- The cost to fix the damage after an occurrence
- The cost to prevent the threat, including by remediation of vulnerabilities
- The goal or intent of the probable attacker

1.3.4 Risk Classifications:

Describe the sensitivity of the data and the likelihood it may be sought after for malicious purposes.

- **Critical Risk Data (CRD)**:
  - Personal information aggressively sought for unauthorized use by both internal and external parties due to its high direct financial value
  - Compromise of CRD would not only harm individuals, but would result in financial harm to the company from significant penalties, costs to retain customers and employees, as well as harm to brand and reputation
- **High Risk Data (HRD)**:
  - HRD is actively sought for unauthorized use due to its potential direct financial value
  - HRD provides the company with a competitive edge
  - If compromised, it could expose the company to financial harm through loss of opportunity
  - Loss of HRD can cause mistrust leading to the loss of business and may result in legal exposure, regulatory fines and penalties, as well as damage to brand and reputation
- **Moderate Risk Data (MRD)**:
  - Company information that has little tangible value to unauthorized parties; however, the unauthorized use of this non-public information would likely have a negative effect on the company

1.3.5 Data Security Organisation

The Information Security Function depends on the size of the enterprise. May be:

- Dedicated Information Security group within IT
- Chief Information Security officer (CISO) reporting to the CIO or CEO
- Smaller organisations data security is the responsibility of data managers

Dedicated Information Security personnel are most concerned with the technical aspects, such as combating malicious software attacks. Data Management are concerned with regulatory aspects. A standard sharing process should be in place where both IT and Data Management are kept informed of data regulations, data loss threats and data protection requirements.

NIST (National Institute of Standards and Technology) Risk Management Framework:

- All enterprise information must be categorised
- The location of all sensitive information must be known
- Enterprise Data Model is essential

Data managers, IT Developers and cyber security professionals work together to:

- Identify regulated data so that
- Sensitive systems are protected

- User access controls designed to
- Enforce confidentiality, integrity and data regulatory compliance

1.3.6 Security Processes

1.3.6.1 The Four As:

- **Access**: Enable individuals with authorization to access systems in a timely manner
- **Audit**: Review security actions and user activity to ensure compliance with regulations and conformance with company policy and standards
- **Authentication**: Validate users' access
- **Authorization**: Grant individuals privileges to access specific views of data, appropriate to their roles
- **Entitlement**: An Entitlement is the sum-total of all the data elements that are exposed to a user by a single access authorization decision

1.3.6.2 Monitoring

- Systems should include monitoring controls that detect unexpected events
- Real time **active monitoring** for confidential information
- System interruption if an event that does not follow procedure occurs
- **Passive monitoring** takes snapshots at regular intervals

1.3.7 Data Integrity

In security, **data integrity** is the state of being whole, protected from improper deletion, alteration or addition.

1.3.8 Encryption

The process of translating plain text into complex codes to hide privileged information, verify complete transmission or verify the sender's identity. Cannot be read without the decryption key. Four main methods of encryption:

- **Hash:** Uses algorithms to convert data to a mathematical representation. Exact algorithms must be known to reveal the data.
- **Symmetric** (not explained in the DMBOK)
- **Private-key:** Both sender and receiver have the same key
- **Public-key:** Sender uses a public key that is freely available and receiver uses a private key

1.3.9 Obfuscation or Masking

The appearance of the data is changed. Two types of data masking, Persistent and Dynamic:

- **Persistent data masking:** Permanently and irreversibly alters the data. Used for test environments
    - **In-flight persistent masking:** Data is masked when it is moving from source (production) to destination (non-production). Secure as there is no intermediate file
    - **In-place persistent masking:** Source and destination are the same**.** Unmasked data is read, masked then written over the unmasked data
- **Dynamic data masking:** Makes changes to appearance of data to the end user system without changing the underlying data
- **Masking methods:**
    - **Substitution**: Replace characters or whole values with those in a lookup or as a standard pattern. For example, first names can be replaced with random values from a list

- **Shuffling**: Swap data elements of the same type within a record, or swap data elements of one attribute between rows. For example, mixing vendor names among supplier invoices such that the original supplier is replaced with a different valid supplier on an invoice
- **Temporal variance**: Move dates +/– a number of days – small enough to preserve trends, but significant enough to render them non-identifiable
- **Value variance**: Apply a random factor +/– a percent, again small enough to preserve trends, but significant enough to be non-identifiable
- **Nulling or deleting**: Remove data that should not be present in a test system
- **Randomization**: Replace part or all of data elements with either random characters or a series of a single character
- **Encryption**: Convert a recognizably meaningful character stream to an unrecognizable character stream by means of a cipher code. An extreme version of obfuscation in-place
- **Expression masking**: Change all values to the result of an expression. For example, a simple expression would just hard code all values in a large free form database field (that could potentially contain confidential data) to be 'This is a comment field'
- **Key masking**: Designate that the result of the masking algorithm/process must be unique and repeatable because it is being used mask a database key field (or similar). This type of masking is extremely important for testing to maintain integrity around the organization

1.3.10 Network Security Terms:

Refers to data-in-motion across a network

- **Backdoor:** A hidden entry to a computer system bypassing password requirements. Usually left by developers for maintenance. Example is default passwords left unchanged.
- **Bot or Zombie:** A workstation taken over by a Trojan, Virus, Phish or download of an infected file. Bots are remotely controlled to perform malicious tasks
- **Cookie:** Small data file an internet commerce website installs on a computer's hard drive to identify returning visitors and their preferences. Could be used by spyware
- **Firewall**: Software and/or hardware that filters network traffic to protect against unauthorised access or attack
- **Perimeter**: Boundary between organisation's systems and outside. Firewall should be in place here
- **DMZ**: De-militarised Zone. Located between the perimeter firewall and a firewall between it and the internet. Used to pass and temporarily store information moving between organisations
- **Super User Account**: Administrator access to be used in an emergency. Credentials are highly secured and controlled by time, location and user ID
- **Key Logger**: Attack software that captures keystrokes
- **Penetration testing**: An ethical hacker tries to expose vulnerabilities
- **Virtual Private Network (VPN)**: Use the unsecured internet to create an encrypted secure tunnel

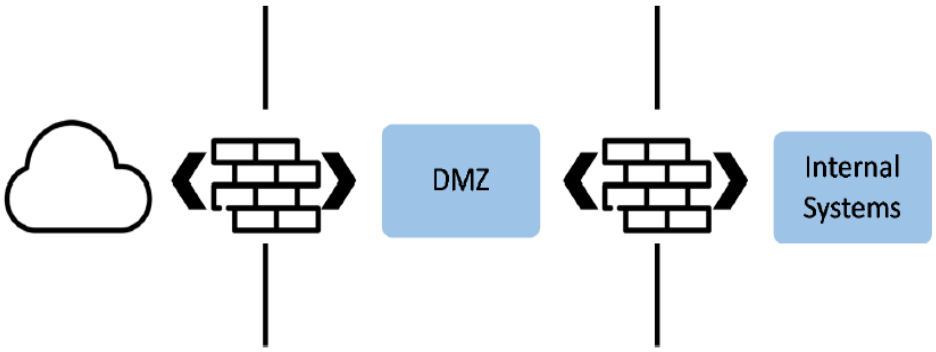


Figure 64 DMZ Example

1.3.11 Types of Data Security

Data security involves not just preventing inappropriate access, but also enabling appropriate access to data. Access to sensitive data is controlled by granting permissions (opt-in). "Least Privilege" is an important security principle.

- **Facility Security:** Locked data centre with access restricted to authorised employees
- **Device Security:** Standards for portable devices
  - Access policies regarding connection using mobile devices
  - Storage of data on portable devices
  - Data wiping and disposal of devices in compliance with records management processes
  - Installation of anti-malware and encryption software
  - Awareness of security vulnerabilities
- **Credential Security:** Each user is assigned User ID and Password to access system
  - **Identity management Systems:** Single sign-on gets user onto many systems
  - **User ID Standards for email systems:** Unique within the system. Usually use the user's name in some way
  - **Password standards:** First line of defence. Should be "strong" and changed every 45-180 days depending on sensitivity of the data, system and enterprise
  - **Multiple Factor Identification:** Additional identification – code to phone, hardware, biometric. Two-factor harder to break than password alone
- **Electronic Communication Security:** Train users not to send confidential information over email or other insecure applications

1.3.12 Types of Data Security Restrictions

Security restrictions are driven by **Confidentiality and Regulation**. The difference is the origin. Confidential restrictions are internal, regulatory restrictions are externally defined.

- **Confidential Data**: Confidential means secret or private and is shared on a "need-to-know" basis. Internally defined. Confidentiality level of a data set depends on the most sensitive item. Typical classification schema:
  - **For General Audiences**: Available to anyone
  - **Internal use only**: Employees or members. May be discussed but not copied outside the organisation
  - **Confidential**: Cannot be shared outside the organisation without a non-disclosure agreement
  - **Restricted Confidential**: Need-to-know

    - **Registered Confidential**: Legal agreement to assume responsibility for the data's secrecy must be signed
- **Regulated Data:** Regulatory categories are assigned according to laws. Shared on an "allowed-to-know' basis. Externally defined. A single data set may have multiple regulatory categories. It is a good idea to collect many nations personal data privacy laws into a single standard to enforce, achieving international compliance. Sample regulatory families:
    - **Personal Identification Information (PII):** Personally Private Information (PPI). Any information that can identify an individual. EU Privacy Directives
    - **Financially Sensitive Data:** In the US covered by Insider Trading Laws
    - **Medically Sensitive Data / Personal Health Information (PHI):** US covered by HIPAA (Health Information Portability and Accountability Act)
    - **Educational Records:** US covered by FERPA (Family Educational Rights and Privacy Act)
- **Industry or Contract-based Regulations**:
    - **Payment Card Industry Data Security Standard (PCI-DSS):** Any information that can identify an individual with an account at a financial institution
    - **Competitive advantage or trade secrets**: Protected by industry regulations / Intellectual property laws
    - **Contractual restrictions**: An organisation may restrict how information may be shared in their contracts with others

1.3.13 System Security Risks

Identify risks inherent in systems:

- **Abuse of excessive privilege:** Principle of least privilege should be applied. Query-level access control is better, but time consuming to set up
- **Abuse of legitimate privilege:** For unauthorised purposes. Enforce policies for end-point machines using time of day, location and amount of data downloadable
- **Unauthorised Privilege Elevation:** Convert from ordinary user to Administrator using software vulnerabilities. Prevent with Intrusion Prevention Systems (IPS) and query-level access control
- **Service account or shared account abuse:**
    - **Service Accounts**: Batch IDs. Untraceable to a particular user
    - **Shared Accounts:** Generic IDs with one password. Provide ungoverned access
- **Platform intrusion attacks:** Protect databases with Intrusion Protection and Intrusion Detection Systems. Any update patches should be installed immediately
- **SQL Injection vulnerability:** A SQL command is inserted in a web input space**.** Sanitise all inputs before passing to server
- **Default passwords:** Usually supplied by software vendor. Eliminate them
- **Backup data abuse**: Encrypt all database backups and securely manage decryption keys

1.3.14 Hacking/Hacker

A hacker finds unknown pathways in complex computer systems. Can be good or bad:

- White Hat hacker (Western movies the hero always wore a white hat) finds vulnerabilities which are fixed in the patches
- Malicious (Black Hat) hackers intentionally breach systems to steal information or do damage

1.3.15 Social Threats to Security / Phishing

Involves direct communication to trick people to provide confidential information - Social engineering. Phishing is the call or message

1.3.16 Malware

Any malicious software created to damage, change or improperly access a computer or network.

- **Adware**: Spyware that slips into the computer from an internet download. It monitors browsing and buying habits. Not illegal
- **Spyware**: Any program that slips in without consent
- **Trojan Horse**: A malicious program that enters the system embedded in legitimate software
- **Virus:** A program that attaches itself to an executable file, and delivers a destructive payload
- **Worm**: A program built to reproduce and spread across a network by itself. Usually harms networks by consuming bandwidth
- **Malware Sources:**
  - Instant Messaging (IM)
  - Social Networking Sites
  - Spam

2 Activities

2.1 Identify Data Security Requirements

There is no one prescribed way to implement data security. Organisations should design their own security controls.

- **Business Requirements:** Analyse business rules and processes to identify security touch points
- **Regulatory requirements:** Create a central inventory of all data regulations and the data subject areas affected by each regulation

Regulation	Subject Area Affected	Security Policy Links	Controls Implemented

Table 13 Sample Regulation Inventory Table

2.2 Define Data Security Policy

Data security policies describe behaviours determined to be in the best interests of the organisation protecting its data. Must be auditable and audited. Requires **collaboration** between IT Security administrators, Security Architects, Data Governance committees, Data Stewards, internal and external audit teams and the legal department.

2.2.1 Security Policy contents:

- **Enterprise Security Policy:** Global policies
- **IT Security Policy:** Directory structures standards, password policies and identity management framework
- **Data Security Policy:** Categories for individual applications, database roles, user groups and information sensitivity

2.3 Define Data Security Standards

Standards supplement policies and provide detail on how to meet the intention of the policies.

- **Define Confidentiality Levels:** Confidentiality classification is important metadata, guides how users are granted access privileges
- **Define Regulatory Categories:** Regulated Information. Regulations Imply a goal - compliance
- **Define Security Roles:** Two ways to organise
  - **Role Assessment grid:** starting from the data

	Confidentiality Level		
	General Audience	Client Confidential	Restricted Confidential
Not Regulated	Public User Role	Client Manager Role	Restricted Access Role
PII	Marketing Role	Client Marketing Role	HR Role
PCI	Financial Role	Client Financial Role	Restricted Financial Role

Table 14 Role Assignment Grid Example

  - **Role Assessment Hierarchy:** Starting from the User

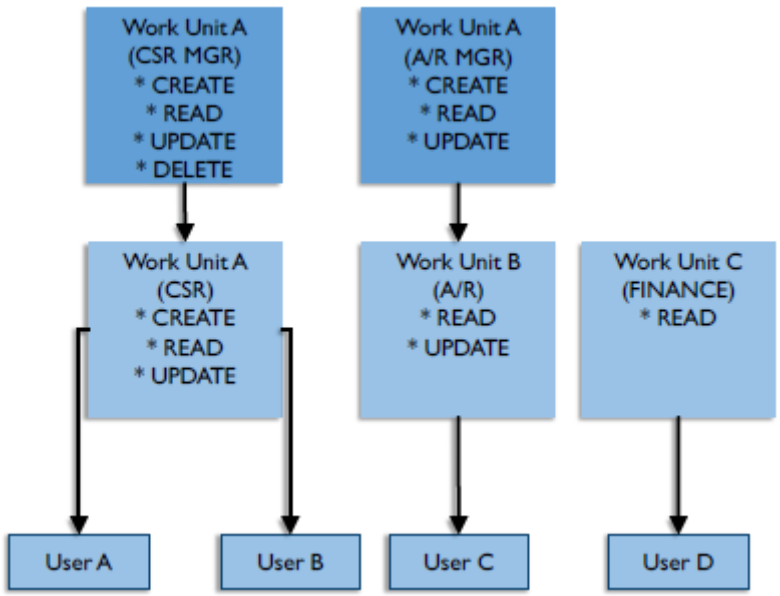


Figure 65 Security Role Hierarchy Example Diagram

- **Assess Current Security Risks:** Identify where sensitive data is stored. Evaluate each system for the following:
  - The sensitivity of data stored of in transit
  - The requirements to protect that data
  - The current security protections in place

  Document the findings as they provide a baseline and may be a requirement of compliance.

- **Implement Controls and Procedures:** Responsibility of security administrators in coordination with data stewards and technical teams. Controls and procedures should at minimum cover:
  - How users gain and lose access to systems
  - How users are assigned to and removed from roles
  - How privilege levels are monitored
  - How requests for access changes are handled and monitored
  - How data is classified according to confidentiality and regulations
  - How data breaches are handled
- **Assign confidentiality levels**
- **Assign regulatory Categories**
- **Manage and maintain data security**:
  - Control Data Availability / Data-centric Security
  - Monitor User Authentication and Access Behaviour:
    - Required for compliance audits
    - Lack of automated recording of sensitive and unusual data base transactions represents serious risks
    - Implement a network based audit appliance to mitigate the risks
- **Manage security policy compliance**
  - Manage regulatory compliance
  - Audit data security and compliance activities

3 Tools

- Anti-Virus Software / Security Software
- Web Page Security: If the web address begins **https://** a security layer is present
- Identity Management Technology
- Intrusion Detection (IDS) and Prevention Software (IPS)
- Firewalls (Prevention)
- Metadata tracking
- Data Masking/Encryption

4 Techniques

- **CRUD Matrix usage:** Create and use data-to-process and data-to-role relationships
- **Immediate Security Patch Deployment:** No one should be able to delay this update
- **Data Security Attributes in Metadata:** Metadata repository is essential
- **Security Needs in Project Requirements:** Identify in the analysis phase
- **Efficient Search for Encrypted Data:** Same encryption method to search for required data to reduce amount of data to decrypt, and therefore reduce risk
- **Document Sanitisation:** Clean the Metadata preventing embedded confidential data being shared

5 Implementation Guidelines

- **Readiness Assessment / Risk Assessment:**
  - **Training:** Training on security initiatives at all levels of the organisation
  - **Consistent policies:** Data security policies should align with enterprise policies
  - **Measure the benefits of security:** Link to organisational activities

    - **Set security requirements for vendors:** Include in SLAs and contracts
    - **Build a sense of urgency:** Emphasise legal, regulatory and contractual requirements to build a sense of urgency
    - **Ongoing Communications**
- **Organisation and Cultural Change**
- **Visibility into User Data Entitlement:** Requires Metadata of classification and the authorisations themselves
- **Data Security in an Outsourced World:** Anything can be outsourced **except liability (Accountability)**
    - Tighter management of control mechanisms
- **Data Security in Cloud Environments:** Data security policies should account for data distributed over these platforms, and should be the same as the rest of the enterprise

6 Data Security Governance

Requires cooperation between IT and business stakeholders, and strong policies and procedures.

6.1 Data Security and Enterprise Architecture

Data security Architecture is a component of Enterprise Architecture that describes how data security is implemented (Enterprise Technology Architecture). Architecture influences:

- Tools used to manage data security
- Data encryption standards and mechanisms
- Access guidelines to external vendors and contractors
- Data transmission protocols over the internet
- Documentation requirements
- Remote access standards
- Security breach access reporting

Security architecture is particularly important for integration of data between:

- Internal systems and business units
- An organisation and its external business partners
- An organisation and regulatory agencies

6.2 Metrics

Frame as positive value percentages

- **Security implementation Metrics:** Maintain a reasonable number of actionable metrics in appropriate categories over time to assure compliance
- **Security Awareness metrics:**
    - Risk assessment findings
    - Risk events and profiles
    - Formal feedback surveys and interviews
    - Incident post-mortems, lessons learned and victim interviews
    - Patching effectiveness audits
- **Data Protection Metrics:**
    - Criticality ranking
    - Annualised loss expectancy
    - Risk of specific data losses

- Risk mapping of data to specific business processes
- Threat assessments
- Vulnerability assessments

- **Security Incident Metrics:**
  - Intrusion attempts detected and prevented
  - Return on investment for security costs using savings from prevented intrusions
- **Confidential Data Proliferation**

Chapter 8
Data Integration and Interoperability

1 Introduction

DII describes processes relating to the movement and consolidation of data within and between data stores, applications and organisations. **Data Integration** consolidates data into consistent forms. **Data Interoperability** is the ability for multiple systems to communicate.

DII provides:

- Data migration and conversion
- Data consolidation into hubs or marts
- Integration of vendor packages into organisation
- Data sharing across applications or organisations
- Distributing data across data stores and data centres
- Archiving data
- Managing data interfaces
- Obtaining and ingesting external data
- Integrating structured and unstructured data
- Providing operational intelligence and management decision support

DII is dependent on these other areas of data management:

- **Data Governance:** The transformation rules and message structures
- **Data Architecture**: Designing solutions
- **Data Security**: Ensuring solutions protect the security of data
- **Metadata:**
    - Tracking technical inventory of data (persistent, virtual and in motion)
    - Business meaning of data
    - Business rules for transforming data
    - Operational history and lineage of the data
- **Data Storage and Operations:** Managing the physical instantiation of the solutions
- **Data Modelling and Design:** Designing the physical and virtual data structures, and messages passing information between applications and organisations

DII is critically important **to Data Warehousing and Business Intelligence** as well as **Reference Data and Master Data Management**.

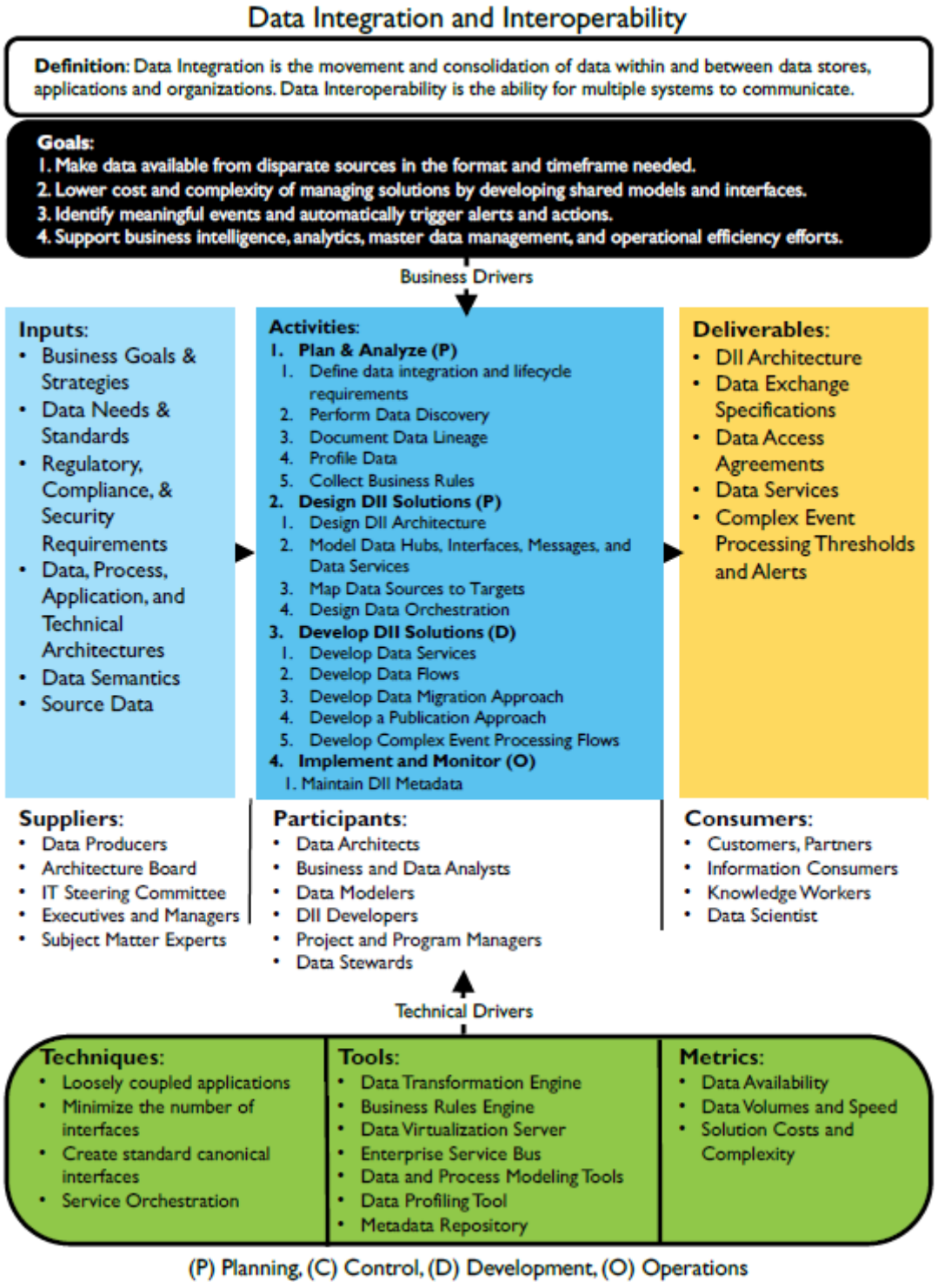

Data Integration and Interoperability
Definition: Data Integration is the movement and consolidation of data within and between data stores, applications and organizations. Data Interoperability is the ability for multiple systems to communicate.
Goals:
1. Make data available from disparate sources in the format and timeframe needed.
2. Lower cost and complexity of managing solutions by developing shared models and interfaces.
3. Identify meaningful events and automatically trigger alerts and actions.
4. Support business intelligence, analytics, master data management, and operational efficiency efforts.
Business Drivers
Inputs:
• Business Goals & Strategies
• Data Needs & Standards
• Regulatory, Compliance, & Security Requirements
• Data, Process, Application, and Technical Architectures
• Data Semantics
• Source Data
Activities:
1. Plan & Analyze (P)
1. Define data integration and lifecycle requirements
2. Perform Data Discovery
3. Document Data Lineage
4. Profile Data
5. Collect Business Rules
2. Design DII Solutions (P)
1. Design DII Architecture
2. Model Data Hubs, Interfaces, Messages, and Data Services
3. Map Data Sources to Targets
4. Design Data Orchestration
3. Develop DII Solutions (D)
1. Develop Data Services
2. Develop Data Flows
3. Develop Data Migration Approach
4. Develop a Publication Approach
5. Develop Complex Event Processing Flows
4. Implement and Monitor (O)
1. Maintain DII Metadata
Deliverables:
• DII Architecture
• Data Exchange Specifications
• Data Access Agreements
• Data Services
• Complex Event Processing Thresholds and Alerts
Suppliers:
• Data Producers
• Architecture Board
• IT Steering Committee
• Executives and Managers
• Subject Matter Experts
Participants:
• Data Architects
• Business and Data Analysts
• Data Modelers
• DII Developers
• Project and Program Managers
• Data Stewards
Consumers:
• Customers, Partners
• Information Consumers
• Knowledge Workers
• Data Scientist
Technical Drivers
Techniques:
• Loosely coupled applications
• Minimize the number of interfaces
• Create standard canonical interfaces
• Service Orchestration
Tools:
• Data Transformation Engine
• Business Rules Engine
• Data Virtualization Server
• Enterprise Service Bus
• Data and Process Modeling Tools
• Data Profiling Tool
• Metadata Repository
Metrics:
• Data Availability
• Data Volumes and Speed
• Solution Costs and Complexity
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 66 Context Diagram: Data Integration and Interoperability

1.1 Business Drivers

- The need to manage data movement efficiently
- Purchased applications come with its own data stores that must integrate with the other data store in the organisation
- An enterprise view of data integration is more cost effective than point to point solutions
- Data hubs such as data warehouses and Master Data solutions

- Managing the cost of support by using standard tools and reducing the complexity of interface management
- DII supports the organisation's ability to comply with data handling standards and regulations

1.2 Goals and Principles

Goals:
1. Make data available from disparate sources in the format and timeframe needed.
2. Lower cost and complexity of managing solutions by developing shared models and interfaces.
3. Identify meaningful events and automatically trigger alerts and actions.
4. Support business intelligence, analytics, master data management, and operational efficiency efforts.

When implementing DII follow these principles:

- Design should take an enterprise perspective (for future extensibility), but implement iteratively and incrementally
- Balance local data needs with enterprise data needs, including support and maintenance
- Ensure business accountability for DII design and activity

1.3 Essential concepts

1.3.1 Extract, Transform and Load (ETL)

The essential steps in moving data around:

- **Extract:** Select required data and extract it from its source, and stage it physically or in memory
- **Transform:** Make the data compatible with the structure of the target store. May be done in batch or real-time:
  - **Format changes:** Technical format e.g., EBCDIC to ASCII
  - **Structure changes:** e.g., Denormalised to normalised
  - **Semantic conversion:** Conversion of values to maintain consistent semantic representation
  - **De-duping:** If rules require unique values, scan target and remove duplicate rows
  - **Re-ordering:** Change to order of the file data elements to fit a pattern
- **Load:** Physically store the result of the transformation in the target system

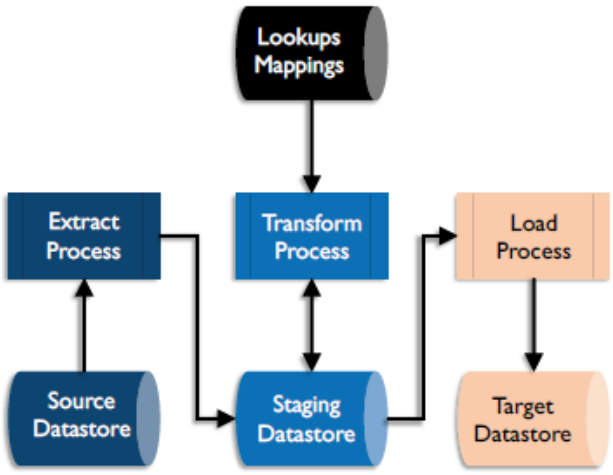


Figure 67 ETL Process Flow

- **ELT (Extract, Load and Transform):** Used if the target system has more transformation capability. also allows data to be instantiated on the target as raw data

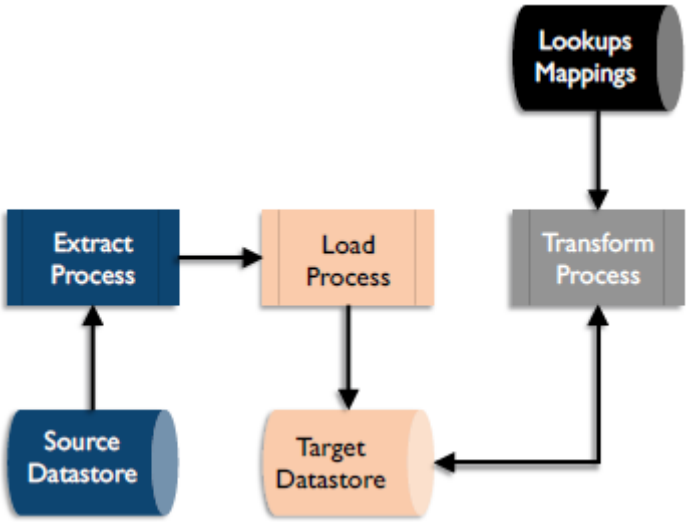


Figure 68 ELT Process Flow

- **Mapping:** A synonym for transformation, the process of developing a lookup matrix from source to target structures, and the result of that process

1.3.2 Latency

The time difference between when data is generated in the source system and when it is available in the target system. Can be high (batch), low (event driven) to very low (real-time synchronous).

- **Batch:** Data moving in clumps or files periodically. Called a **batch** or **ETL**
  - **Snapshot:** Full set of data at a point in time
  - **Delta:** Data that has changed values since the last time data was sent
  - High latency
  - Used for data conversions, migrations, archiving and extracting from and loading data warehouses and marts
- **Change Data Capture:** Filtering data to include only the data that has changed in a given timeframe (the delta) and passes it on to data consumers
- **Near-real-time and Event Driven:** Data is processed in smaller sets spread across the day, or when and event such as an update occurs. Lower latency than batch
- **Asynchronous:** The source does not wait for the target to acknowledge before continuing processing. The target need not be available
- **Real-time, Synchronous:** No time delay or other differences between source and target are acceptable. The executing process waits for confirmation before executing its next transaction
- **Low Latency or Streaming:** Extra hardware costs. Need extremely fast transfer of large amounts of data over large distances

1.3.3 Replication

Applications maintain exact copies of data sets on multiple physical locations. Better response times for international users. Use DBMS replication utilities. Not recommended if changes to the data occur at more than one site.

1.3.4 Archiving

ETL functions can transport and possibly transform infrequently used data to a cheaper storage solution.

1.3.5 Enterprise Message Format/ Canonical Model

A canonical model is the common model used by an organisation to standardise the format in which data will be shared. Transformations need only be done to and from the canonical model.

- Hub-and-spoke
- All systems interact with central information hub
- Data is transformed based on the enterprise message format of the organisation
- Reduces transformations as each system only needs to transform data to and from the central canonical model
- Reduces complexity of DII in the enterprise
- Lowers cost of support
- Complex to develop
- Justified in managing more than 3 systems and critical for more than 100

1.3.6 Interaction Models

Describe ways to make connections between systems:

- **Point-to-point:**
  - Systems pass data directly to each other
  - Suitable in small systems
  - Can impact processing
  - Many interfaces to manage
  - Multiple interfaces can lead to inconsistent data
- **Hub-and-spoke:**
  - Consolidates shared data in a central hub
  - Examples are Data Warehouses, Data Marts. Operational Data Stores and Master Data Management Hubs
  - Easy to add more systems
  - Enterprise Service Buses (ESB) – near real-time sharing of data where the hub is a virtual canonical model
- **Publish-subscribe:** Systems push out (publish) data and other systems pull data in (subscribe)

1.3.7 DII Architecture Concepts

- **Application coupling:**
  - **Tight coupling**: Synchronous interface, one waits for the other to respond
  - **Loose coupling**: Preferred interface design as data is passed without waiting for a response and without causing both systems to be unavailable if one is unavailable. e.g., Service Oriented Architecture using an Enterprise Service Bus

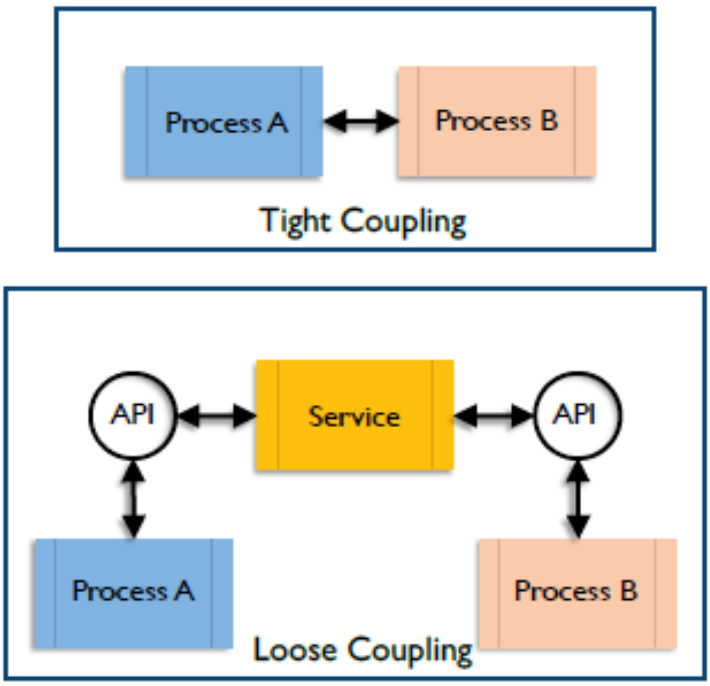


Figure 69 Application Coupling

- **Orchestration and Process Controls:**
  - **Orchestration:** How multiple processes are organised and executed in a system. All systems handling messages must be able to manage the order of those processes
  - **Process Controls:** The components that ensure shipment, delivery, extraction and loading of data is complete. Include:
    - Database activity logs
    - Batch job logs
    - Alerts
    - Exception logs
    - Job dependence charts with remediation options, standard responses
    - Job clock information – length of jobs and computing window time
- **Enterprise Application Integration (EAI):** Software modules only interact with each other through APIs
- **Enterprise Service Bus (ESB):** An intermediary system, passing messages between other systems

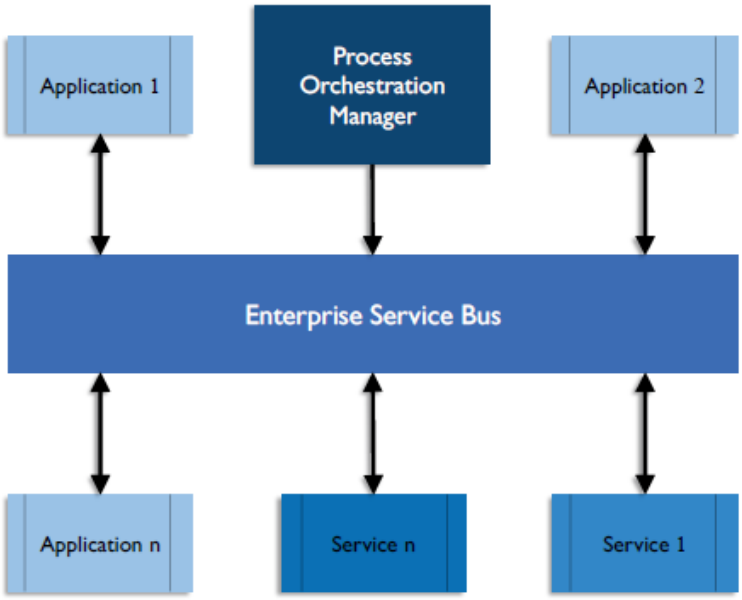


Figure 70 Enterprise Service Bus

The CDM does not comprise of all data within the organisation, but merely the data that is shared on the integration layer. Data that is internal to a domain would therefore not be included during the design of a Canonical Data Model.

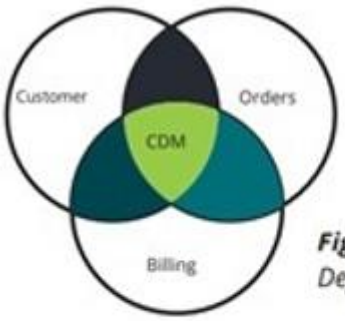


Figure 1: *Defining the CDM*

A business has a single Canonical Data Model

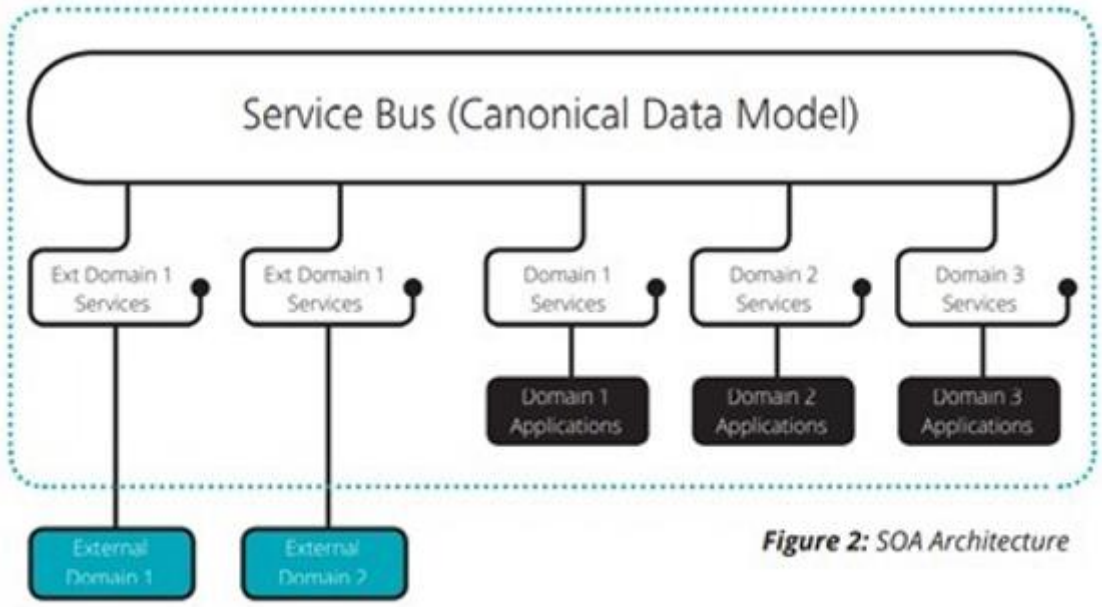


Figure 2: *SOA Architecture*

SOA Architecture (Author Unknown) NOTE in this example CDM stands for Canonical Data Model

- **Service-Oriented Architecture (SOA):** The functionality of providing data or updating data is provided through well-defined service calls between applications enabling application independence. The goal of SOA is well defined interaction between self-contained, independent software modules. Usually implemented through APIs
- **Complex Event Processing (CEP):** Tracks and analyses data about events from multiple sources to predict meaningful events (threats, opportunities) to predict behaviour or automatically trigger a response. Big Data. Requires technologies that support ultra-low latency
- **Data Federation and Virtualisation:** Provides access to a combination of individual data stores. Virtualisation enables distributed databases to be accessed as a single database
- **Data-as-a-Service (DaaS):** Data licensed from a vendor and provided on demand
- **Cloud-based Integration:** Also called Integration Platform-as-a-service (IPaaS) and are usually run as SaaS applications at the data centres of vendors

1.3.8 Data Exchange Standards

Data Exchange Standards are formal rules for the structure of data elements. ISO or any common model used by an organisation or data exchange group. It simplifies data interoperability, lowers support costs and increases understanding of the data

2 Data Integration Activities

Data Integration activities follow a development lifecycle (plan, design, development, testing and implementation).

2.1 Plan and Analyse

2.1.1 Define data integration and lifecycle requirements:

- Data and technology required to meet business objectives
- Laws and restrictions on data contents
- Defined by Business Analysts, data stewards and various architects
- Creates and uncovers valuable Metadata – manage throughout lifecycle
- Perform Data Discovery:
- Identify potential sources of data for the integration effort
- High level assessment of data quality
- Maintain the inventory of organisational data in a Metadata repository
- Plan for acquiring and integrating external data
- Document Data Lineage:
- How data under analysis is acquired or created by the organisation, where it moves, how it is changed, how it is used by the organisation for analytics, decision making or triggering
- Profile Data:
- Data profiling reveals actual data structure and contents
- Assess the quality of the data
- Can reveal differences from what is assumed leading to early intervention
- Basic profiling involves analysis of:
    - Data format (data structures and inferred from actual data)
    - Data population (NULLS, blanks and defaults)
    - How data values correspond to a set of valid values
    - Patterns and relationships internal to the data set
    - Relationships to other data sets

- The requirement to profile data must be balanced with an organisation's privacy and security rules (see Chapter 13)
- Collect Business Rules: rules harvesting or mining
- **A Business Rule is a statement that defines or constrains and aspect of business processing**
- Four categories:
    - Definition of business terms
    - facts relating terms to each other
    - Constraints or action assertions
    - Derivations
- Use business rules to support DII to:
    - Assess data in potential source and target data sets
    - Direct the flow of data in the organisation
    - Direct when to automatically trigger events and alerts

2.2 Design Data Integration Solutions

2.2.1 Design Data Integration Architecture

DI solutions should be specified at both enterprise and individual solution level. Enterprise standards save time as planning has been done and group licences and resource sharing and reusing DII components saves money.

- **Solution architecture indicates**:
    - Techniques and technologies to be used
    - Inventory of involved data structures
    - Orchestration and frequency of data flow
    - Regulatory and security concerns and remediation
    - Operating concerns around backup and recovery, archiving and retention
- **Select Interaction model:** Hub-and-spoke, point-to-point or publish-subscribe
- **Design data services or exchange patterns:** Start with industry standards or existing patterns

2.2.2 Model Data Hubs, Interfaces, Messages and Data Services

Model persistent and transient datatypes.

2.2.3 Map Data Sources to Targets

Specify the rules for transforming data from one location and format to another. For each attribute mapped specify:

- Technical format of source and target
- Transformations required for intermediate staging points
- How each attribute in the final or intermediate data store will be populated
- Whether data values need to be transformed
- Calculations required

2.2.4 Design Data Orchestration

Pattern of data flows from start to finish, including intermediate steps.

- **Batch**: Frequency of data movement and transformation is usually coded into a scheduler
- **Real-time**: Usually triggered by an event such as an update

2.3 Develop Data Integration Solutions

- **Develop Data Services:** Can be tools or vendor suites
- **Develop data flows:**
  - ETL flows developed in a tool such as a scheduler for batch which manages the order, frequency and dependency of executing the data pieces
  - Real-time: Monitor for events that trigger services to acquire, transform or publish data
- **Develop data migration approach:** Moving data needs proper analysis and testing
- **Develop a publication approach:** Push changed data using common message definitions (canonical model) and notify the recipients
- **Develop complex event processing flows:**
  - Preparing historical data needed for the predictive model
  - Pre-populate predictive model and identify meaningful events
  - Executing triggered action in response to a prediction

2.4 Implement and Monitor

- Activate the data services that have been developed and tested
- Real-time data processing needs real-time monitoring for issues, human or automated
- Monitor and service at the level of the most demanding target application or consumer
- **Maintain DII Metadata:**
  - Manage Metadata uncovered during the development of DII solutions
  - Document all data structures involved
  - ETL vendors have Metadata repositories
  - Service-Oriented Architecture (SOA) registry is particularly important for DII

3 Tools

- **Data Transformation Engine/ETL Tool:** Primary tool that supports operation and design
- **Data Virtualisation Server:** Perform data extract, transform and integrate virtually. A data warehouse is often input to a data virtualisation server
- **Enterprise Service Bus:** Middleware to support near real-time messaging between heterogeneous sources in the same enterprise
- **Business Rules Engine:** Allows non-technical users to manage business rules implemented by software
- **Data and Process Modelling Tools:** Used to design target and intermediate data structures
- **Data Profiling tool:** Use a tool to analyse large amounts of data to determine if it meets requirements
- **Metadata Repository:** Tools have Metadata repositories which store the rules for transformation, lineage and processing, as well as the instructions for scheduled processes and triggers

4 Techniques

Described in Essential Concepts.

5 Implementation Guidelines

5.1 Readiness Assessment / Risk Assessment

As all organisations already have DII in place assess for readiness/risk around tool implementation or interoperability. An enterprise data integration solution supports the movement of data between many applications and organisations.

Working DII solutions shouldn't be replaced. Focus on where none exists. Additional use of data integration adds to the investment in a data warehouse to Master Data Management hub.

It is necessary to sponsor the implementation of an enterprise data integration program by a high level of authority over solution design and technology purchase, to prevent local data integration solutions from developing. It may be perceived the cost of these is less than the enterprise wide solution.

Don't become too focussed on the tool and lose focus on the business needs.

5.2 Organisation and Cultural change

- Local teams understand data in their applications
- Central teams know tools and techniques
- Enterprise solutions should be overseen by a Centre of Excellence
- Although technical, data integration solutions must be based on deep business knowledge to successfully deliver value
- Modelling and data analysis should be done by business resources
- Canonical message model (consistent standard for how data is shared in the organisation) development requires technical and business resources
- Business SMEs review transformation mapping design and changes

6 DII Governance

Business is responsible for:

- Decisions about the design of data messages, data models and data transformation rules
- Defining the rules for loading and transforming data
- Approve changes to these rules (which are captured as Metadata for cross enterprise analysis)
- Identifying and verifying predictive models and defining the actions they trigger

Governance controls to support trust that the DII will perform as promised:

- Determine what events trigger governance reviews (exceptions or critical events)
- Map each trigger to reviews that engage with governance bodies
- Event triggers may be part of SDLC at Stage Gates or part of User Stories

Controls come from governance-driven management routines such as mandated review of models, Metadata audits, gating of deliverables or required approval of changes to the transformation rules.

Include real-time operational data integration solutions in SLAs and Business Continuity/Disaster Recovery plans on the same tier as the most critical system to which they provide data.

Policies need to be established to ensure the organisation benefits from an enterprise approach to DII.

6.1 Data Sharing agreements

A data sharing agreement or memorandum of understanding (MOU) must be put in place before developing interfaces, and be approved by business data stewards, which stipulates:

- Responsibilities
- Acceptable use of data to be exchanged
- Anticipated use and access to the data
- Restriction on use
- Expected service levels
- Required system up times and response times

6.2 DII and Data Lineage

Governance to ensure that knowledge of data origins and movement is documented. Data lineage must be managed as it is critical Metadata.

Compliance standards require an organisation be able to describe where its data originated, and how it has changed as it moves through systems.

Impact analysis when making changes to data structures, data flows or data processing requires forward and backward data lineage.

6.3 Data Integration metrics

To measure scale and benefits of DII solutions:

- Data Availability (of data requested)
- Data volumes and speed
    - Volumes of data transported and transformed
    - Volumes of data analysed
    - Speed of transmission
    - Latency between update and availability
    - Latency between event and triggered action
    - Time to availability of new data sources
- Solution costs and complexity
    - Cost of developing and managing solutions
    - Ease of acquiring new data
    - Complexity of solutions and operations
    - Number of systems using data integration solutions

Chapter 9
Document and Content Management

1 Introduction

Document and Content Management entails controlling the capture, storage, access and use of data and information **stored predominantly outside relational databases**. Records include both paper documents and ESI (Electronically Stored Information). Ensuring security and quality requires governance, reliable architecture, and well-managed Metadata.

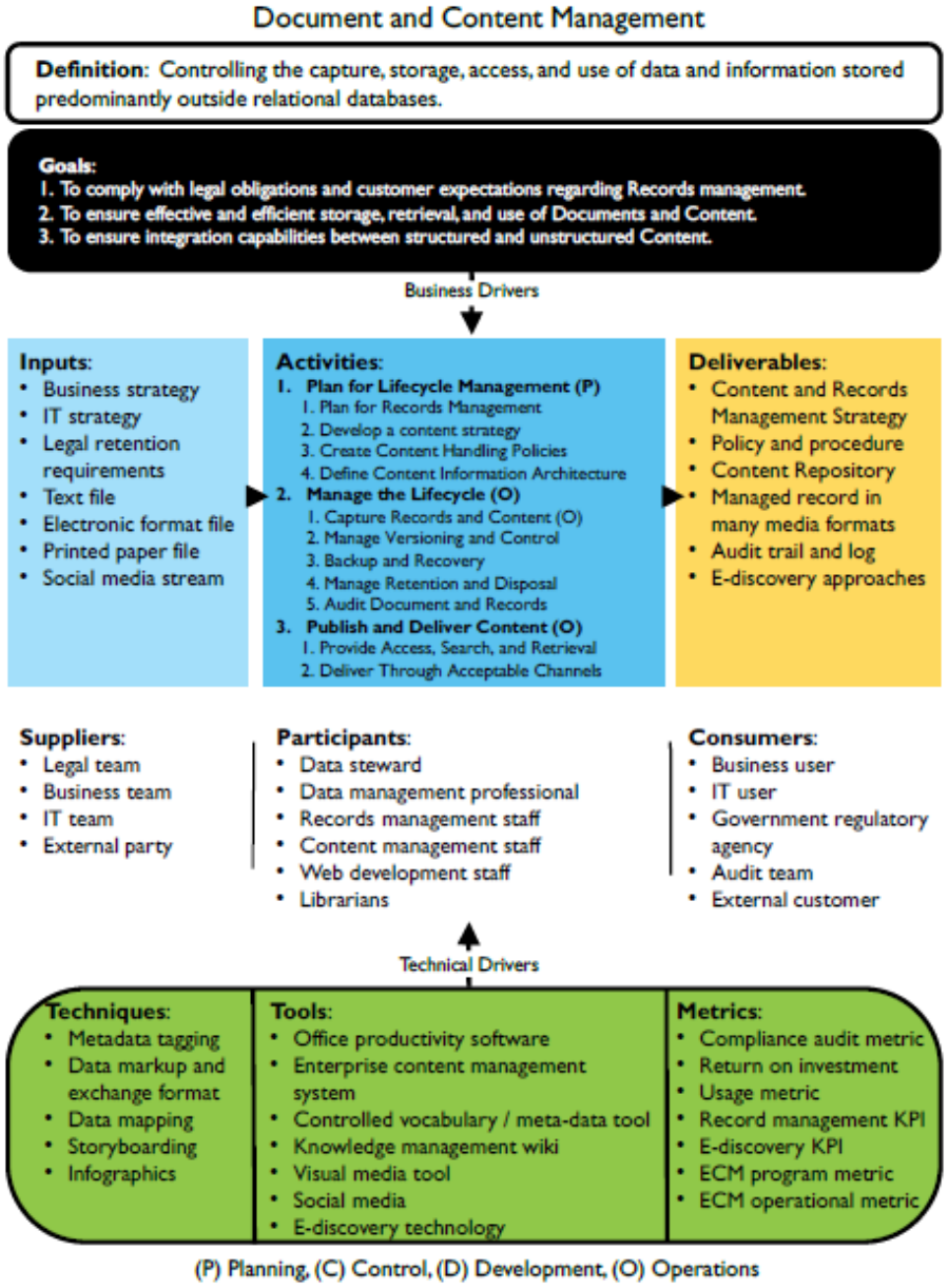

Document and Content Management
Definition: Controlling the capture, storage, access, and use of data and information stored predominantly outside relational databases.
Goals:
1. To comply with legal obligations and customer expectations regarding Records management.
2. To ensure effective and efficient storage, retrieval, and use of Documents and Content.
3. To ensure integration capabilities between structured and unstructured Content.
Business Drivers
Inputs:
• Business strategy
• IT strategy
• Legal retention requirements
• Text file
• Electronic format file
• Printed paper file
• Social media stream
Activities:
1. Plan for Lifecycle Management (P)
1. Plan for Records Management
2. Develop a content strategy
3. Create Content Handling Policies
4. Define Content Information Architecture
2. Manage the Lifecycle (O)
1. Capture Records and Content (O)
2. Manage Versioning and Control
3. Backup and Recovery
4. Manage Retention and Disposal
5. Audit Document and Records
3. Publish and Deliver Content (O)
1. Provide Access, Search, and Retrieval
2. Deliver Through Acceptable Channels
Deliverables:
• Content and Records Management Strategy
• Policy and procedure
• Content Repository
• Managed record in many media formats
• Audit trail and log
• E-discovery approaches
Suppliers:
• Legal team
• Business team
• IT team
• External party
Participants:
• Data steward
• Data management professional
• Records management staff
• Content management staff
• Web development staff
• Librarians
Consumers:
• Business user
• IT user
• Government regulatory agency
• Audit team
• External customer
Technical Drivers
Techniques:
• Metadata tagging
• Data markup and exchange format
• Data mapping
• Storyboarding
• Infographics
Tools:
• Office productivity software
• Enterprise content management system
• Controlled vocabulary / meta-data tool
• Knowledge management wiki
• Visual media tool
• Social media
• E-discovery technology
Metrics:
• Compliance audit metric
• Return on investment
• Usage metric
• Record management KPI
• E-discovery KPI
• ECM program metric
• ECM operational metric
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 71 Context Diagram: Document and Content

1.1 Business Drivers

- **Regulatory compliance**
- **The ability to respond to litigation:** Respond efficiently and consistently
- **e-discovery:** The process of finding electronic records that might serve as evidence in litigation. Includes email, chats, websites, raw application data and Metadata. Huge volume of ESI. Proactive management is essential
- **Business continuity requirements**
- **Efficiency**: Use technology to streamline processes, manage workflow, eliminate repetitive manual tasks and enable collaboration
- **Customer satisfaction**: Well organised and searchable websites

1.2 Goals and Principles

Goals:
1. To comply with legal obligations and customer expectations regarding Records management.
2. To ensure effective and efficient storage, retrieval, and use of Documents and Content.
3. To ensure integration capabilities between structured and unstructured Content.

Goals of implementing best practices around Document and Content Management:

- Ensuring effective and efficient retrieval and use of data in unstructured formats
- Ensuring integration capability between structured and unstructured data
- Complying with legal obligations and customer expectations

Management of Document and Content follow these guiding principles:

- Everyone in the organisation must create, use, retrieve and dispose of records in accordance with the established policies and procedures. Train everyone
- Experts in the handling of records and content should be fully engaged in policy and planning

In 2009 ARMA International published GARP (Generally Acceptable Recordkeeping Principles®) how business records must be maintained: Details on ARMA website

- **Principle of Accountability:** Senior executive assigns appropriate individuals, adopts policies and processes and ensures auditability
- **Principle of Integrity:** Guarantee of authenticity and reliability
- **Principle of Protection:** Personal information is protected
- **Principle of Compliance:** Program should comply with laws, binding authorities and the policies of the organisation
- **Principle of Availability:** Information is managed in a manner that ensures timely, efficient and accurate retrieval of information
- **Principle of Retention:** Retain information for an appropriate time taking legal and regulatory requirements into consideration
- **Principle of Disposition:** Provide secure and appropriate disposition of information in accordance with operational, legal, regulatory and fiscal requirements
- **Principle of Transparency:** Policies, processes, activities, governance program documented and understood by all staff and appropriate interested parties

1.3 Essential Concepts

1.3.1 Content

Document is to content what a bucket is to water: a container or file. **Content** refers to the data and information inside the file, document or website. Content has a lifecycle and can become an official **record**. Official records are treated differently to other content.

- **Content Management:**
  - The processes, techniques and technologies for organising, categorising and structuring information resources so that they can be stored, published and reused
  - Important in websites
  - Indexing based on keywords from taxonomies
  - Enterprise Content Management (ECM) for entire enterprise
- **Content Metadata:** Essential for managing unstructured data and "Big Data"
  - **Format:** Often dictates the method to access the data
  - **Search-ability:** Whether search tools already exist
  - **Self-documentation:** Whether the Metadata self-documenting as in file systems
  - **Existing patterns:** Whether existing methods can be adopted or adapted
  - **Content subjects:** The things people are likely to be looking for
  - **Requirements:** Need for thoroughness and detail in retrieval means a content tagging tool is necessary
- **Content Modelling:** The process of converting logical content concepts into content types, attributes and data types with relationships using Metadata Management and Data Modelling techniques. Two levels of content modelling:
  - Information product level e.g., creating a website
  - Component level with detail the elements in the information product
- **Content Delivery Methods:** Content needs to be modular, structured, reusable and device independent. Convert to XML to enable reuse over many channels. Types of delivery systems:
  - **Push:** Users choose type of content to be delivered on a schedule
  - **Pull:** Users pull content through the Internet
  - **Interactive:** e.g., electronic point of sale. High volumes of real-time data. Options include EAI, CDC, Data Integration and EII

1.3.2 Controlled Vocabularies

A defined list of explicitly allowed terms to index, categorise, tag, sort and retrieve content through browsing and sorting, and to systematically organise content. Controlled vocabularies should be aligned with the entity names and definitions in the Enterprise Conceptual Model. May be considered **Reference Data** and **Metadata**.

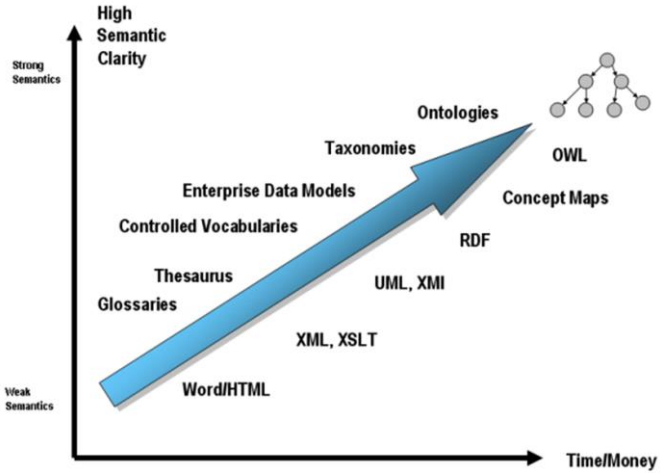
- **Vocabulary Management:** The function of defining, sourcing, importing and maintaining any given vocabulary.
  - ANSI/NISO Z39.19-2005 definition: Way to improve the effectiveness of information storage and retrieval systems, web navigation systems and other environments that seek to both identify and locate desired content via a description using language
  - Purpose of vocabulary control is consistency in the description of content objects to facilitate retrieval
- **Vocabulary Views and Micro-Controlled Vocabulary:**

    - A **vocabulary view** is a subset of a large standard vocabulary relevant to a group of users e.g., Finance only needs financial terms
    - A **micro-controlled vocabulary** is a vocabulary view containing highly specialised terms not present in the general vocabulary
- **Term and Pick Lists:** Lists in applications that do not contain relationships, such as menus or drop-downs. They control the domain of values
- **Term Management:** One or more words designating a concept (ANSI/NISO Z39.19-2005). Terms should be specified and managed through governance processes. Three types of term relationships:
    - **Equivalent term relationship:** Leads to one or more terms to use
    - **Hierarchical relationship:** Depicts broader (general) to narrower (specific) or whole-part relationships
    - **Related term relationship:** A term that is associatively but not hierarchically linked to another term in a controlled vocabulary
- **Synonym rings and Authority lists:**
    - **Synonym ring:** A set of terms with roughly equivalent meaning. A search on one reveals content on the others
    - **Authority list:** A controlled vocabulary of descriptive terms designed to facilitate retrieval of information within a specific domain. Terms have different authority levels, some are preferred, and some are variants. Must have a designated manager
- **Taxonomies:** Any classification or controlled vocabulary. A naming structure containing a controlled vocabulary used for outlining topics and enabling navigation and search systems. Can be a data model.
    - **Flat taxonomy:** No relationships among the set of controlled categories. A list
    - **Hierarchical taxonomy:** Tree structure where nodes are related by a rule. Going up expands the category and going down refines.
    - **Polyhierarchy:** Tree structure with more than one rule. Multiple parents. Paths are complex
    - **Facet taxonomy:** Star structure. Each node is associated to a central node e.g., Metadata where each attribute is a facet of a content object
    - **Network taxonomy:** Both hierarchical and facet structures, e.g., a thesaurus
- **Classification Schemes and Tagging:**
    - Classification schemes are codes that represent controlled vocabulary, e.g., Dewey Decimal System for libraries
    - **Folksonomy:** For online content and terms and names are obtained though social tagging
- **Thesauri:** A Thesaurus is a controlled vocabulary used for content retrieval. Provides information about each term and its relationship to other terms
- **Ontology:**
    - A type of taxonomy that represents a set of concepts and their relationships **within a domain**
    - Primary knowledge representation in the **Semantic Web**
    - Developed using ontology languages such as OWL (Web Ontology Language) and RDFS (Resource Description Framework Schema)
    - Describe classes (concepts), Individuals (instances) attributes, relations and events
    - Two **differences** between taxonomy and ontology:
        - **Taxonomy:**

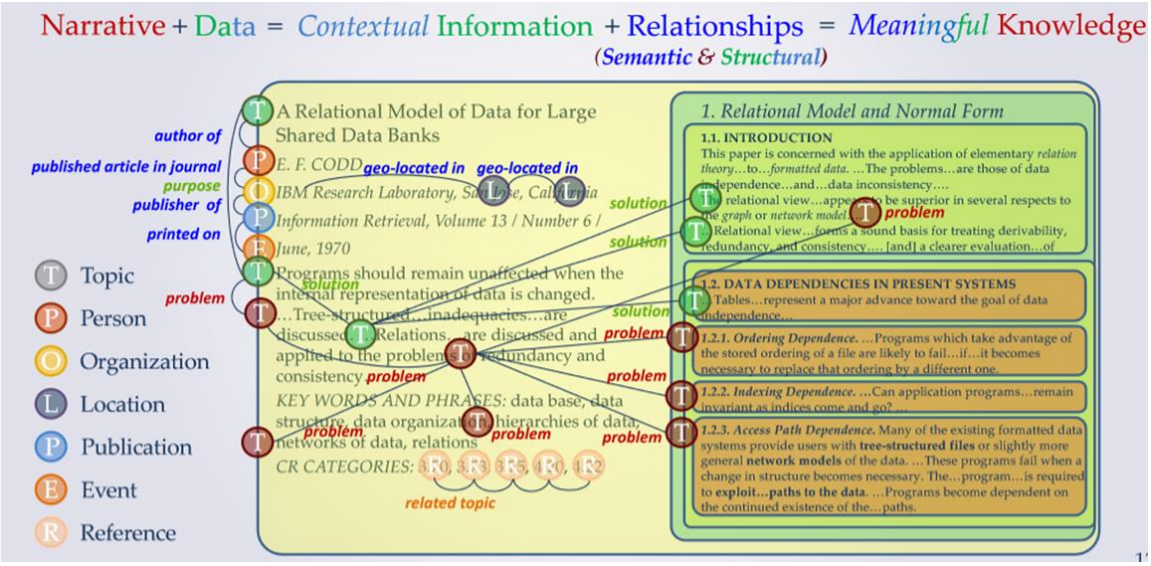
- Like a data model
- Data content classifications for a given concept area
- Entity, Attribute and valid values are defined
- **Closed-world assumption**
  - Only what is **known** is defined

▪ **Ontology:**

- Content concepts are mixed and identified through Metadata
- **Open-world assumption**
  - Possible relationships are inferred



Relationship between Technology and Meaning, Depending on Investment (Author Unknown)

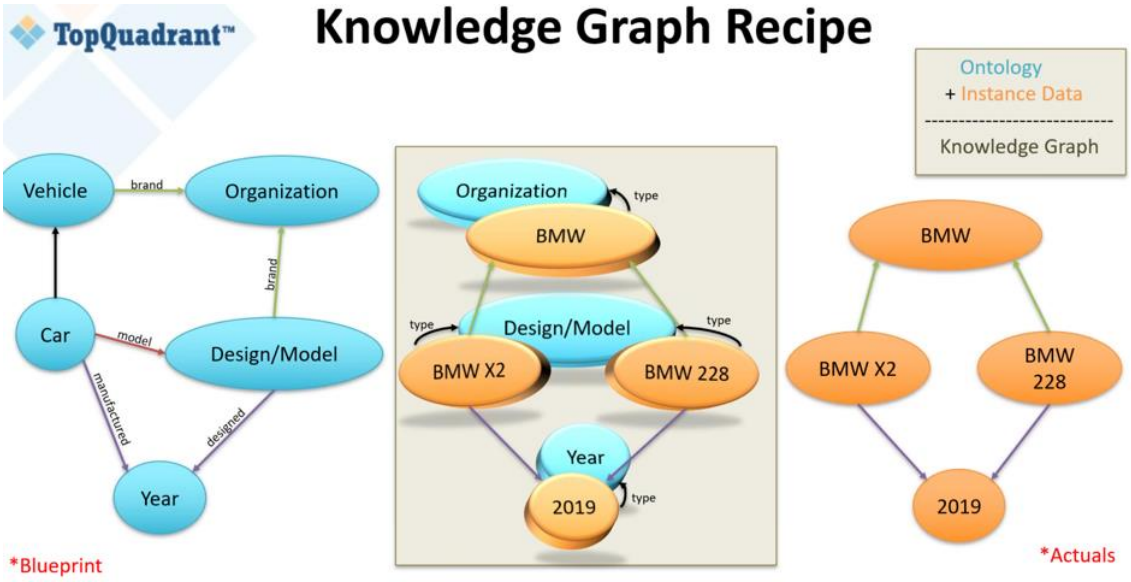

Narrative + Data = Contextual Information + Relationships = Meaningful Knowledge
(Semantic & Structural)
A Relational Model of Data for Large Shared Data Banks
E. F. CODD
IBM Research Laboratory, San Jose, California
Information Retrieval, Volume 13 / Number 6 / June, 1970
Programs should remain unaffected when the internal representation of data is changed.
...Tree-structured...inadequacies...are discussed. Relations...are discussed and applied to the problem of redundancy and consistency.
KEY WORDS AND PHRASES: data base, data structure, data organization, hierarchies of data, networks of data, relations
CR CATEGORIES:
related topic
author of
published article in journal
purpose
publisher of
printed on
geo-located in
solution
problem
Topic
Person
Organization
Location
Publication
Event
Reference
1. Relational Model and Normal Form
1.1. INTRODUCTION
This paper is concerned with the application of elementary relation theory...to...formatted data. ...The problems...are those of data independence...and...data inconsistency....
Relational view...forms a sound basis for treating derivability, redundancy, and consistency.... [and] a clearer evaluation...of
1.2. DATA DEPENDENCIES IN PRESENT SYSTEMS
Tables...represent a major advance toward the goal of data independence...
1.2.1. Ordering Dependence. ...Programs which take advantage of the stored ordering of a file are likely to fail...if...it becomes necessary to replace that ordering by a different one.
1.2.2. Indexing Dependence. ...Can application programs...remain invariant as indices come and go? ...
1.2.3. Access Path Dependence. Many of the existing formatted data systems provide users with tree-structured files or slightly more general network models of the data. ...These programs fail when a change in structure becomes necessary. The...program...is required to exploit...paths to the data. ...Programs become dependent on the continued existence of the...paths.

This image shows how machines can extract information from a document using language constructs: Subject, Predicate, Object (Text Mining)

Legend on this diagram:

Narrative (red) + Data (green/light blue) = Contextual (light blue) Information (green) + Relationships (dark blue) = Meaningful (dark blue/light blue/green) knowledge (red). Semantic (dark blue) & Structure (green/light blue)

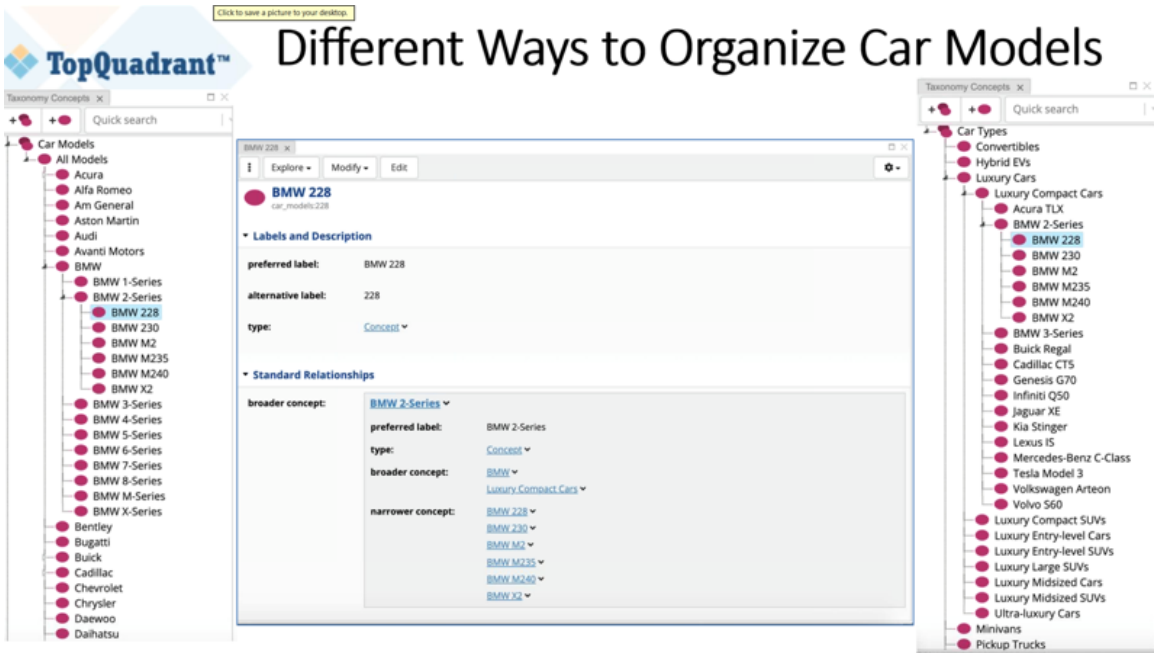
Can ETL into graph database to become a word map.



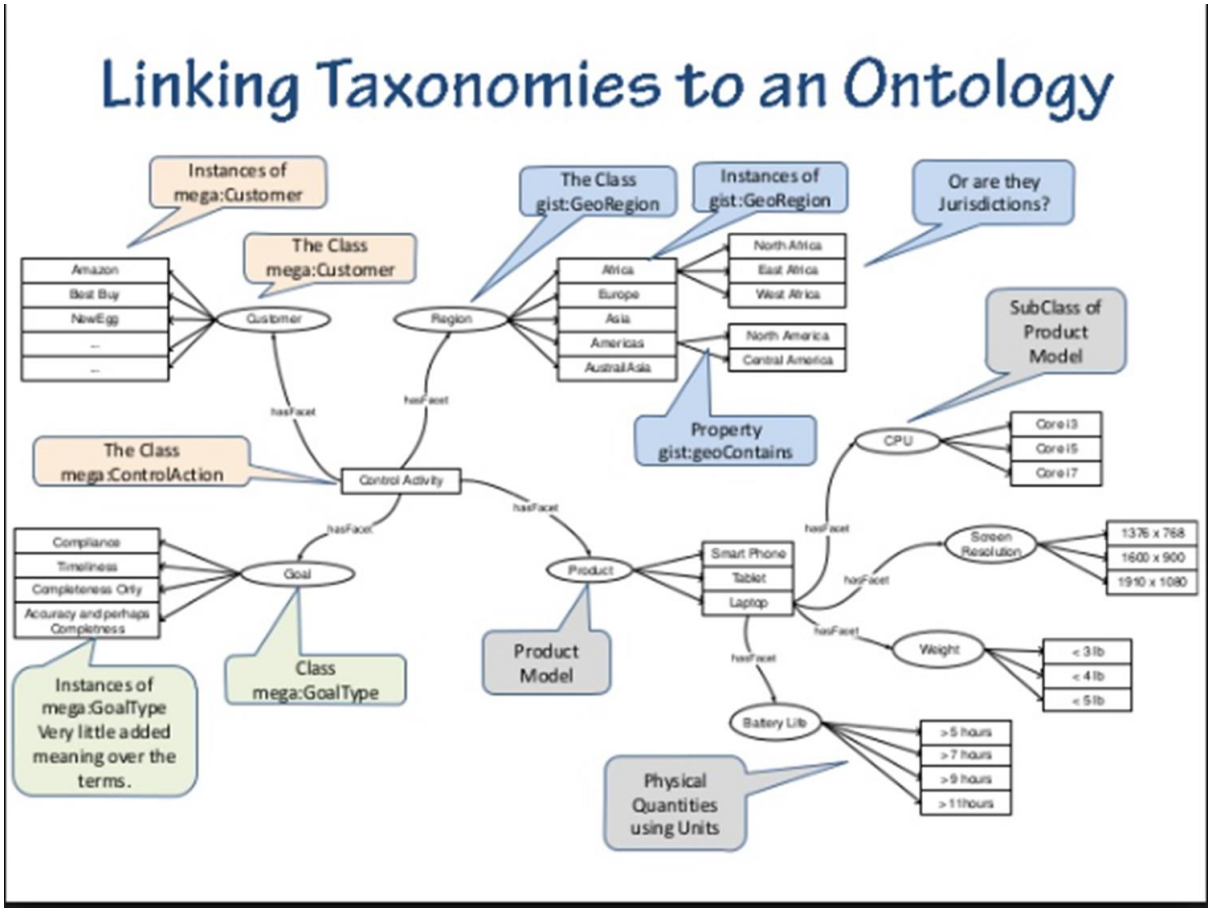
The blue diagram on the left is an ontology – a data model (schema) with classes, properties and rules, a precise blueprint for data. Orange diagram on the far right is the BMW instance data (actuals) which can be extracted from unstructured / raw data. Overlaid they become a Knowledge Graph. **Ontology (blue) + Instance Data (orange) = the knowledge graph**

An ontology stands for the study of **WHAT EXISTS** and gives us more semantic clarity about what the document is talking about.

A knowledge graph / semantic network represents a network of real-world entities—i.e. objects, events, situations, or concepts—and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph." This is a process of tagging, Once we can map the contents of unstructured data to our ontology we get better clarity.



Example of a **Polyhierarchy**. There are 2 ways to get to BMW 228. Multiple paths therefore multiple parents



An example of a **Network Taxonomy**. Note the combination of Facet and Hierarchical. **Is a** – classification **Hierarchy**, and **Has** a – associations between notes, grouping or structure (laptop) – **Ontology**

1.3.3 Documents and Records

Documents are electronic or paper objects which communicate information and knowledge.

Records are a subset of documents which provide evidence of actions taken and decisions made in accordance with procedures. Can be created manually or automatically by equipment.

- **Document Management:** Processes, techniques and procedures for controlling and organising documents and records throughout their lifecycle
  - **Inventory:** Identification of existing and newly created documents and records
  - **Policy:** Creation, approval and enforcement of documents and records policies
  - **Classification**
  - **Storage:** Short and long term physical storage
  - **Retrieval and circulation:** Allow access in accordance with policies, security and legal standards
  - **Preservation and disposal:** Archiving and destroying documents and records according to organisational needs, statutes and regulations

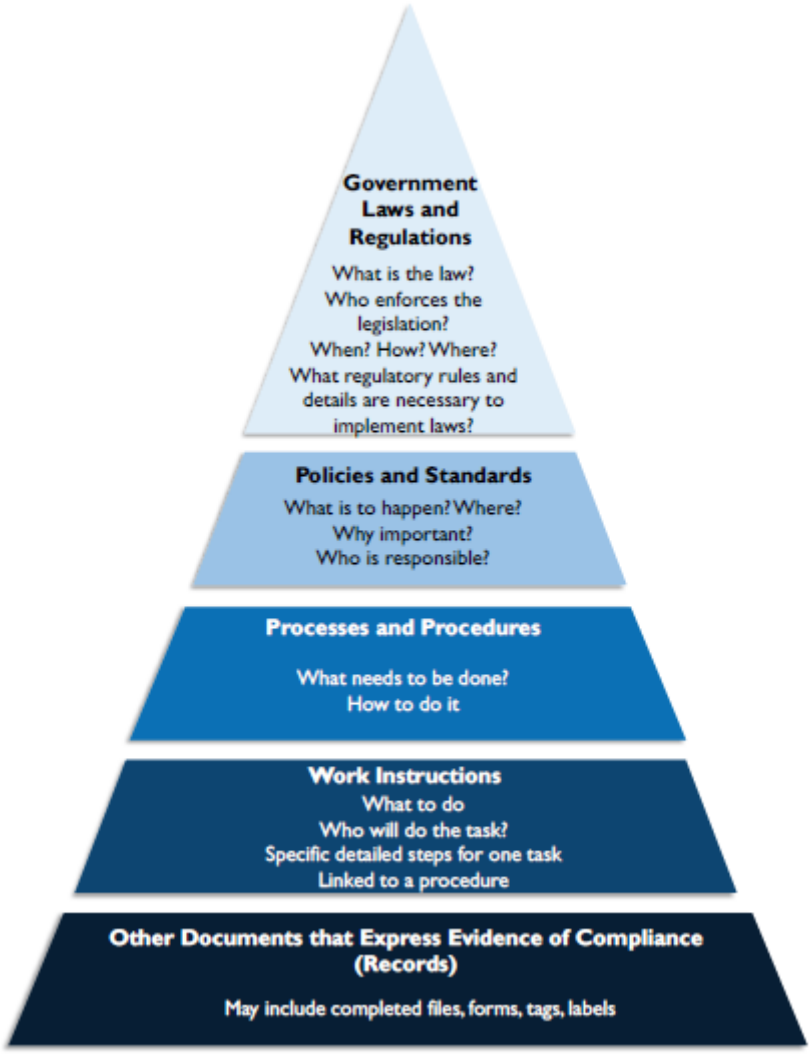

Government Laws and Regulations
What is the law?
Who enforces the legislation?
When? How? Where?
What regulatory rules and details are necessary to implement laws?
Policies and Standards
What is to happen? Where?
Why important?
Who is responsible?
Processes and Procedures
What needs to be done?
How to do it
Work Instructions
What to do
Who will do the task?
Specific detailed steps for one task
Linked to a procedure
Other Documents that Express Evidence of Compliance (Records)
May include completed files, forms, tags, labels

Figure 72 Document Hierarchy based on ISO 9001 – 4.2

- **Records Management:**
  - Managing records through the full lifecycle:
    - Record creation or receipt
    - Processing
    - Distribution
    - Organisation
    - Retrieval
    - Disposition
  - Records can be physical, electronic, web content, documents on all kinds of media, data in any kind of database
  - A **Vital Record**: A type of record required by an organisation to resume operations during and after a disaster
  - Trustworthy records have signatures for regulatory compliance and integrity
  - Characteristics of well-prepared records:
    - **Content**: Accurate, complete, truthful
    - **Context:** Collect Metadata
    - **Timeliness:** Created promptly after the event

- **Permanency:** Records may never be changed
- **Structure:** Content should be clear and legible

- **Digital Asset Management:** Document management for rich media documents (videos, logos, photographs etc.)

1.3.4 Data Map

A data map is an inventory of all ESI (Electronically Stored Information) data sources, applications and IT environments.

1.3.5 E-discovery

Discovery is the pre-trial phase of a lawsuit where both parties request information from each other to find the facts of the case. Rules were amended in 2006 to include ESI (Electronically Stored Information). Electronic documents have Metadata which plays a part in evidence.

- **Legal requirements**:
  - e-discovery
  - data and records retention practices
  - Legal hold notification (LHN) process
  - Legally defensible disposition practices

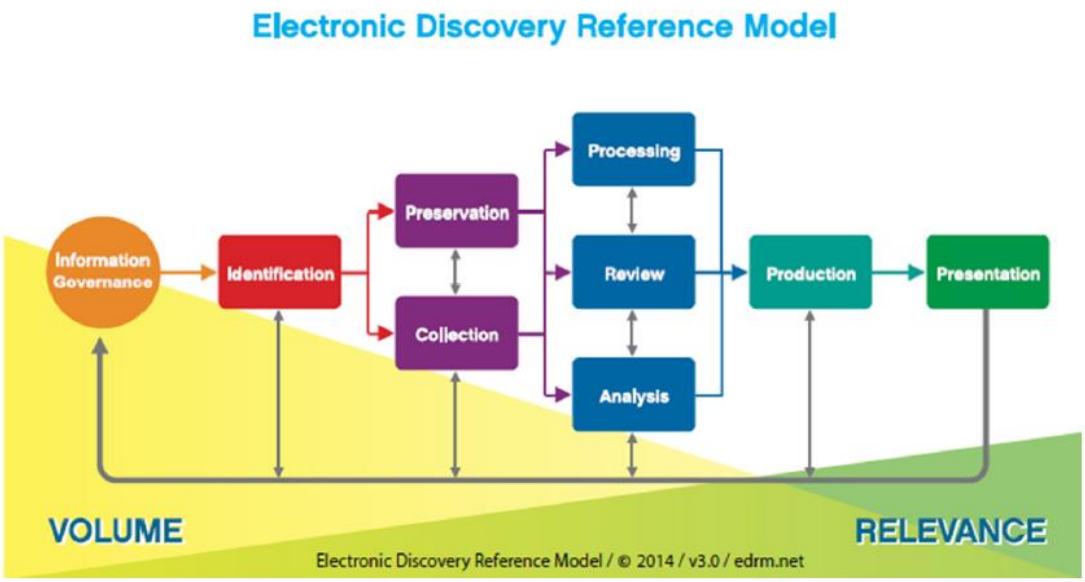


Figure 73 Electronic Discovery Reference Model

EDRM is a standards and guidelines organisation for e-discovery. The EDRM model assumes information governance is in place. Eight e-discovery phases:

- **Identification:**
  - **Early Case Assessment:** Legal case assessed for descriptive information or Metadata
  - **Early Data Assessment:** Type and location of data relevant to the case is assessed
    - Identify policies for retention or destruction of relevant data so ESI can be preserved
    - Interviews with involved personnel to get pertinent information
    - Involved personnel need to understand the case background, legal hold and their roles
- **Preservation:** Identified potentially relevant data is put on hold to avoid destruction

- **Collection:** Acquisition and transfer information to legal counsel in a legally defensible manner
- **Processing:** Data de-duplicated, searched and analysed to determine which moves to review
- **Review:** Documents identified to be presented on request
- **Analysis:** Understand the content to formulate a legal strategy
- **Production:** Information turned over to opposing counsel based on agreed to specifications
- **Presentation:** Displaying the ESI at depositions, hearings and trials

1.3.6 Information Architecture

Information architecture is the process of creating structure for a body of information or content. Identifies links and relationships between documents and content. Components:

- Controlled vocabularies
- Taxonomies and ontologies
- Navigation maps
- Metadata maps
- Search functionality specifications
- Use cases
- User flows

The information architecture and the content strategy together describe WHAT content will be managed in the system. Design phases describe HOW the strategy will be implemented.

1.3.7 Search Engine

A search engine is software that searches for information based on terms and retrieves websites that have those terms within their content. e.g., Google.

1.3.8 Semantic Model

Semantic modelling is a type of modelling that describes a network of concepts and their relationships, and allows users to ask questions of the system in a non-technical way.

1.3.9 Semantic Search

A semantic search engine can use artificial intelligence to identify query matches based on keywords and their context. BI and analytics tool users often have semantics search requirements.

1.3.10 Unstructured Data

Also non-tabular or semi-structured data. About 80% of all stored data is outside relational databases. Include word processing documents, emails, social media, chats, flat files, spreadsheets, XML files, transactional messages, reports, graphics, digital images, microfiche and video and audio recordings.

There is no data model to aid understanding. No tags. Not organised in rows and columns.

The fundamental principles of Data Management apply. Unstructured data requires data governance, architecture, security, Metadata and data quality.

1.3.11 Workflow

Manage content management through a workflow that ensures content is created on schedule and receives proper approvals. Should be automated through use of a content management system (CMS) to provide version control.

2 Activities

2.1 Plan for Lifecycle Management

Planning for document management:

- For the document's lifecycle
- Developing classification/indexing systems and taxonomies
- Creating policy specifically for records

Identify an organisational unit responsible for managing documents and records.

2.1.1 Plan for Records Management

Define what constitutes a record. The defining team for a functional area should include SMEs and records system people. Decisions:

- Where to store and archive
- Account for paper records as well as electronic unstructured and structured data

2.1.2 Develop a Content Strategy

Defines how content will be prioritised, organised and accessed. Identify Content drivers (the reasons content is needed).

- Inventory of current state
- Gap assessment

2.1.3 Create Content Handling Policies

Policies help employees understand and comply with the requirements for document and records management.

Most document management programs have policies relating to:

- Scope and compliance with audits
- Identification and protection of vital records
- Purpose and schedule for retaining records
- How to respond to information hold orders (in the event of a lawsuit holding records beyond expiry of retention)
- Requirements for onsite and offsite storage of records
- Use and maintenance of hard drive and shared network drives
- Email management, addressed from content management perspective
- Proper destruction methods
- **Social Media Policies:** Does a social media post from an employee conducting business become a record?
- **Device Access Policies:** Content and records management functions need to work with user driven IT (BYOD bring-your-own-devices)
- **Handling Sensitive Data:** Data Security and Data Governance establish confidentiality schemes and identify confidential and restricted assets. People who assemble content must apply these classifications
- **Responding to Litigation:** Prepare for the possibility of litigation through proactive e-discovery. Create an inventory of data sources and their associated risks

2.1.4 Define Content Information Architecture

Users need to submit search to a system which contains unstructured data in a form understandable to that system. The inventory of structured and unstructured documents needs to be described or indexed for quick retrieval.

2.2 Manage the Lifecycle

2.2.1 Capture Records and Content

Capturing data:

- Electronic content is in the form to be stored in electronic repositories
- Paper: scan and upload using an electronic signature
- Tag/index with the appropriate Metadata necessary for retrieval:
  - Document identifier
  - Date and time of capture
  - Title and author(s)

2.2.2 Manage Versioning and Control

ANSI Standard 859 has three levels based on criticality of the data:

- **Formal control:** Formal change control processes through change control authority
- **Revision control:** Notify stakeholders and increment versions when a change is required.
- **Custody control:** Least formal, requires safe storage and means of retrieval.

Data Asset	Formal	Revision	Custody
Action item lists		x	
Agendas			x
Audit findings		x	x
Budgets	x		
DD 250s			x
Final Proposal			x
Financial data and reports	x	x	x
Human Resources data		x	
Meeting minutes			x
Meeting notices and attendance lists		x	x
Project plans (including data management and configuration management plans)	x		
Proposal (in process)		x	
Schedules	x		
Statements of Work	x		
Trade studies		x	
Training material	x	x	
Working papers			x

Table 15 Levels of Control for Documents per ANSI-859

ANSI 859 recommends taking the following into account when determining control level for an asset:

- Cost of providing and updating the asset
- Project impact of changes in terms of cost
- Other consequences of change to the enterprise or project
- Need to reuse or earlier versions of the asset
- Maintenance of a history of change

2.2.3 Backup and Recovery

Include the documents/record management system in the organisation's **corporate backup and recovery plan** and **Business Continuity Plan**. A **vital records program** allows the business to conduct business during a disaster and resume normal business afterwards. Vital records must be identified and plans for their protection and recovery must be developed and maintained.

2.2.4 Manage Retention and Disposal

A retentions and disposition policy defines (in accordance with legal and regulatory requirements):

- The timeframes during which documents for operational, legal, fiscal or historical value must be maintained
- When inactive documents can be transferred to a secondary storage facility
- Processes for compliance
- Methods and schedules for disposition of documents

Records managers and information asset owners ensure **privacy and data protections requirements** and prevent identity theft.

Software considerations: May require a certain version of the operating system to read. An upgrade can make documents unreadable or inaccessible.

Non-value-added information must be identified and removed as it wastes resources and can be discoverable in the event of litigation. This is often not prioritised because:

- Policies are not adequate
- One person's non-value-added information is another's valued information
- Inability to see future needs for current non-value-added records
- There is no buy-in for records management
- Inability to decide which to delete
- Perceived cost of making the decision and removing physical and electronic records
- Electronic space is cheap. Buying more space is easier than archival or removal processes

2.2.5 Audit Documents/Records

Document / Records Management Component	Sample Audit Measure
Inventory	Each location in the inventory is uniquely identified.
Storage	Storage areas for physical documents / records have adequate space to accommodate growth.
Reliability and Accuracy	Spot checks are executed to confirm that the documents / records are an adequate reflection of what has been created or received.
Classification and Indexing Schemes	Metadata and document file plans are well described.
Access and Retrieval	End users find and retrieve critical information easily.
Retention Processes	The retention schedule is structured in a logical way either by department, functional or major organizational functions.
Disposition Methods	Documents / records are disposed of as recommended.
Security and Confidentiality	Breaches of document / record confidentiality and loss of documents / records are recorded as security incidents and managed appropriately.
Organizational understanding of documents / records management	Appropriate training is provided to stakeholders and staff as to the roles and responsibilities related to document / records management.

Table 16 Sample Audit Measures

Steps of an audit:

- Define organisational drivers and stakeholders – the why
- Gather data on the process (the how), determine what to measure and what tools to use
- Report the outcomes
- Develop an action plan of next steps and timeframes

2.3 Publish and Deliver Content

- **Provide access, search and retrieval** once it is described by Metadata/key word tagging and classified within the appropriate information content architecture
- **Deliver through acceptable channels**: Users want to consume data on the device of their choice meaning the format may change. It may be difficult to bring it back to the original format e.g., HTML to originally structured

3 Tools

3.1 Enterprise Content Management Systems

An ECM consists of a platform of core components or applications which can be in-house or in the cloud.

Reports can be delivered through common tools or a document management system interface allowing search, view, download and print. Report management is the ability to add, change or delete reports organised in folders. Reports retention can be set for automatic purge or archive.

3.1.1 Document Management

A **document management system** is an application used to track and store electronic documents and electronic images of paper documents. Provides storage, versioning, security, Metadata management, content indexing and retrieval.

Once a document is created within the document management system it is indexed with keywords so that it can be found. Metadata (Extracted automatically or added manually) is stored for each document. Bibliographic records of documents are descriptive structured data in MARC (Machine-Readable Cataloguing) standard format.

A document repository enables check-in and check-out features, versioning, collaboration, comparison, archiving, status state, migration from one storage media to another and disposition.

May have a module that supports different types of workflows: Manual, rules-based or dynamic rules based on content.

Document management systems have a rights module to allocate permissions.

Specialised document management:

- **Digital asset management:** Document management systems may include management of audio, video and digital photographs
- **Image processing:** Captures, transforms and manages images of paper and electronic documents. Uses:
    - **Scanning:** Digitising
    - **Optical Character Recognition (OCR):** Mechanical or electronic conversion of digitised print or handwriting to form recognised by software

    - **Intelligence character recognition (ICR):** More intelligent, convert cursive handwriting
    - **Form processing:** Capture of printed forms. System recognises the format
    - **Other digitised images:** Binary image file formats:
        - **vector**: Mathematical formulae to create graphics. Can be resized without compromising resolution. File formats: .eps, .ai and .pdf
        - **Raster**: bitmap of coloured pixels. .jpeg, .gif, .png, .tiff
- **Records management system:**
    - Automation of retention and disposition
    - E-discovery support
    - Long term archiving to comply with legal and regulatory requirements
    - Vital records program for critical records

3.1.2 Content Management System

A **content management system** is used to collect, index and retrieve content, storing it either as components or whole documents while maintaining links between the components. It maintains content throughout the lifecycle.

3.1.3 Content and Document workflow

Workflow tools support business processes, route content and documents, assign work tasks, track status and create audit trails.

3.2 Collaboration tools

For teams.

3.3 Controlled vocabulary and Metadata tools

Office productivity software, Metadata repositories, BI tools and document and content management systems.

3.4 Standard Markup and exchange formats

To allow computer systems to process unstructured data it must be reformatted.

- **XML:** Extensible Markup Language
    - Uses Metadata to describe content, structure and business rules of a document
    - Translates structure of the data into a document for data exchange
    - XML tags data elements to identify the meaning of data
    - Older markup methods are HTML and SGML
    - Need for XML-capable content management:
        - Integrating structured and unstructured data in a relational database
        - XML can build enterprise or corporate portals
        - XMP provides labelling and identification of unstructured data
- **JSON:** JavaScript Object Notation
    - Text format is language independent and easy to parse
    - Two structures:
        - Objects: a collection of unordered name/value pairs
        - An array: an ordered list of values
    - Preferred format for web and NoSQL databases
- **RDF and related W3C specifications:**
    - **Resource Description Framework** is the standard for web data exchange

- Stored in a **triplestore:** Subject (Resource) – predicate (property name) – Object (property value) which is a graph database. Queried using SPARQL.
- URI (Uniform Resource Identifier) describes subject-predicate-object
- URL (Uniform Resource Locator) is a form of URI
- Semantic Web needs access to both data and relationships between data sets (Linked Data). This is provided by URIs
- RDF uses XML as its encoding syntax
- SKOS (Simple Knowledge Organisation System) is an RDF vocabulary
- OWL (W3C Web Ontology Language) is a vocabulary extension of RDF used for publishing and sharing OWL documents on the web
- Big Data: If data is accessible using the RDF Triples model, SPARQL query language can be used to find patterns without predefining a schema

- **Schema.org:**
  - Open source vocabulary collection
  - Labelling content with open source semantic markup
  - Enables semantic search engines to index content
  - **Web crawlers** match content with a search query

3.5 E-Discovery technology

Large volumes of data. Technology Assisted Review (TAR) is a process where a team can review selected documents and mark them relevant or not. This is input to a predictive coding engine that reviews and sorts the remaining documents according to relevancy.

4 Techniques

- Litigation Response Playbook for e-discovery
- Litigation Response Data Map to enable the organisation to respond to litigation more effectively

5 Implementation Guidelines

5.1 Readiness Assessment / Risk Assessment

Identify areas where content management improvement is needed. A Data Management Maturity Assessment Model can help.

Specific ECM critical success factors:

- Content audit and classification for existing content
- Appropriate information architecture
- Support of the content lifecycle
- Definitions of appropriate Metadata tags
- The ability to customise functions of the ECM solution

5.1.1 Records Management Maturity

ARMA Generally Accepted Recordkeeping Principles guides policy and practice for Records Management. ARMA Information Governance Maturity Model assesses the organisation's recordkeeping program:

- **Level 1 Sub-Standard:** Information governance and recordkeeping concerns not addressed or just minimally

- **Level 2 In Development:** Developing recognition information governance and recordkeeping can have an impact
- **Level 3 Essential:** Minimum requirements to meet legal and regulatory requirements
- **Level 4 Proactive:** A proactive information governance
- **Level 5 Transformational**: Information governance is integrated into the corporate infrastructure and business processes

Analyse gaps and risks and their impact on the organisation. If an organisation does not inventory records it is already at risk.

5.1.2 E-Discovery Assessment

A readiness assessment examines and identifies improvements to the litigation process. A mature process will be documented, defensible and auditable and will specify:

- Clear roles and responsibilities
- preservation protocols
- data collection methodologies
- disclosure processes

5.2 Organisation and Cultural Change

ECM can lead to more tasks, such as scanning and defining Metadata.

Management of content and records may be siloed. Training may be necessary to enforce policies across the enterprise.

Content and records management should be elevated to a high priority activity organisationally, especially in highly regulated industries where the RIM function needs to be closely aligned with Corporate legal and e-discovery functions.

6 Document and Content Governance

6.1 Information Governance Frameworks

Documents, records and other unstructured content poses risk to an organisation. Drivers to managing the risk include:

- Legal and regulatory compliance
- Defensible disposition on records
- Pro-active preparation for e-discovery
- Security of sensitive information
- Management of risks associated with email and Big Data

Principles of successful Information Governance programs:

- ARMA GARP® principles
- Assign executive responsibility for accountability
- Educate employees on information governance responsibilities
- Classify information correctly
- Ensure authenticity and integrity of information
- Determine that the official record is electronic unless specified directly
- Develop policies for alignment of business systems and third-parties to information governance standards

- Store, manage and make accessible, monitor and audit approve enterprise repositories and systems for records and content
- Secure confidential or personally identifiable information
- Control unnecessary growth of information
- Dispose information at the end of its lifecycle
- Comply with requests for information.
- Improve continuously

The Information Governance Reference model (IGRM) shows the relationship of Information Governance with other business functions. The outer ring shows the stakeholders who put policies, standards, processes, tools and infrastructure in place. The centre is a lifecycle diagram.

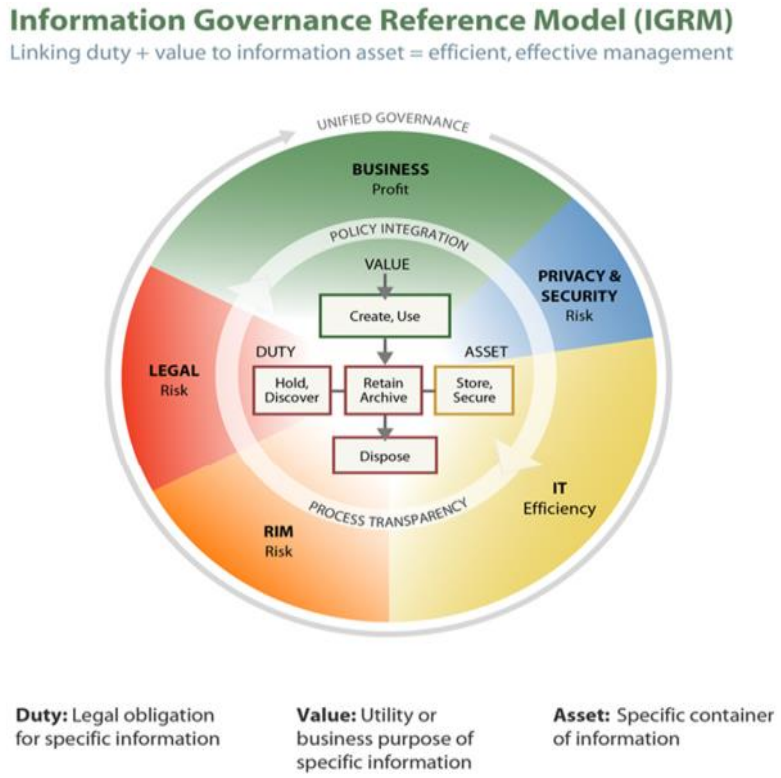


Figure 74 Information Governance Reference Model

6.2 Proliferation of Information

The challenge of governance of unstructured data:

- Unstructured data grows much faster than structured data
- Ownership can be difficult to ascertain
- Difficult to classify as business purpose cannot always be inferred
- Unstructured data without Metadata is a risk as it can be misinterpreted and may be mishandled

6.3 Govern for Quality Content

Requires effective partnership between data stewards and other data management professionals and records managers. Business data stewards can help define web portals, enterprise taxonomies, search engine indexes and content management issues.

Defining quality content requires understanding the context of its production and use:

- **Producers:** Who creates the content and why
- **Consumers:** Who uses the information and for what purpose
- **Timing:** When and how frequently the information is needed
- **Format:** Is a consistent format required? Are there unacceptable formats?
- **Delivery:** How the information is delivered, how it will be accessed and security to prevent inappropriate access

6.4 Metrics

KPIs at strategic and operational levels, especially if they measure lifecycle functions or risks.

6.4.1 Records Management

Strategic level:

- Within regulatory requirements e.g., time taken to meet a requirement
- Governance e.g., compliance with policies

Operational level:

- Within areas of records management resources e.g., operational and capital costs
- Training
- Delivery of daily records management services and operations
- Integration of records management functions with other business functions

Criteria to measure the success of a records management system:

- Percentage of total documents and email per user identified as corporate records
- Percentage of these records put under records control
- Percentage of the total stored records that have proper retention rules applied

Use ARMA's GARP principles and maturity model to guide definition of KPIs.

6.5 E-Discovery

KPIs of e-discovery:

- Cost reduction
- Efficiency of collecting documents ahead of time rather than reactively
- How quickly an organisation can implement a legal hold notification process (LHN)

6.5.1 ECM (Electronic Content Management)

KPIs measure tangible and intangible benefits to the organisation:

- Tangible:
  - Increased productivity
  - Cost reduction
  - Improved information quality
  - Improved compliance
- Intangible:
  - Improved collaboration
  - Simplification of job routines and workflow

Chapter 10
Reference and Master Data

1 Introduction

An organisation and its customers benefit if the data required across business areas, processes and systems is shared, allowing the same customer lists, geographic codes, parts codes etc. to be accessed, to produce a level of consistency.

Systems and data evolve organically, resulting in multiple systems executing similar functions isolated from each other, leading to inconsistencies in data structure and values, and increased costs and risks. Both can be reduced through the management of Reference Data and Master Data.

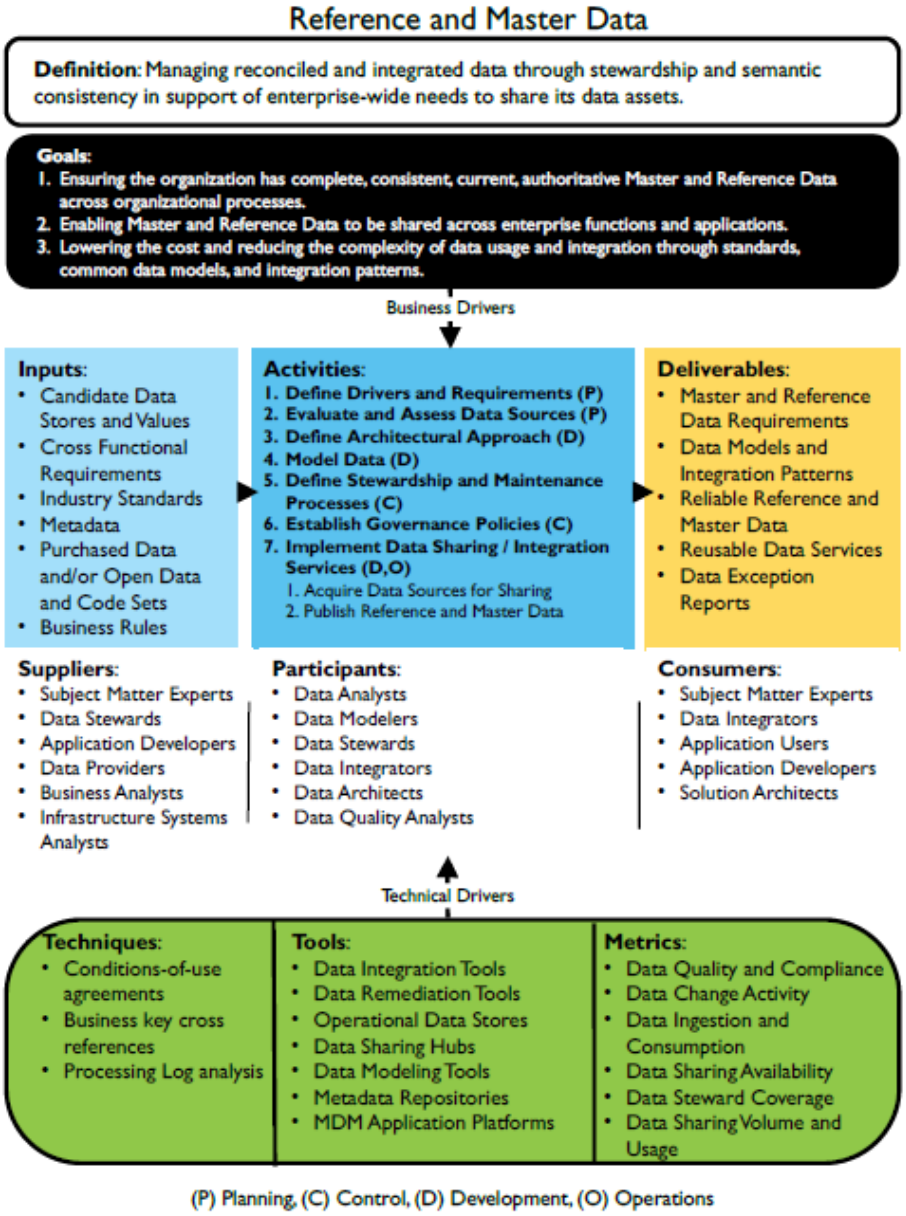

Reference and Master Data
Definition: Managing reconciled and integrated data through stewardship and semantic consistency in support of enterprise-wide needs to share its data assets.
Goals:
1. Ensuring the organization has complete, consistent, current, authoritative Master and Reference Data across organizational processes.
2. Enabling Master and Reference Data to be shared across enterprise functions and applications.
3. Lowering the cost and reducing the complexity of data usage and integration through standards, common data models, and integration patterns.
Business Drivers
Inputs:
• Candidate Data Stores and Values
• Cross Functional Requirements
• Industry Standards
• Metadata
• Purchased Data and/or Open Data and Code Sets
• Business Rules
Activities:
1. Define Drivers and Requirements (P)
2. Evaluate and Assess Data Sources (P)
3. Define Architectural Approach (D)
4. Model Data (D)
5. Define Stewardship and Maintenance Processes (C)
6. Establish Governance Policies (C)
7. Implement Data Sharing / Integration Services (D,O)
1. Acquire Data Sources for Sharing
2. Publish Reference and Master Data
Deliverables:
• Master and Reference Data Requirements
• Data Models and Integration Patterns
• Reliable Reference and Master Data
• Reusable Data Services
• Data Exception Reports
Suppliers:
• Subject Matter Experts
• Data Stewards
• Application Developers
• Data Providers
• Business Analysts
• Infrastructure Systems Analysts
Participants:
• Data Analysts
• Data Modelers
• Data Stewards
• Data Integrators
• Data Architects
• Data Quality Analysts
Consumers:
• Subject Matter Experts
• Data Integrators
• Application Users
• Application Developers
• Solution Architects
Technical Drivers
Techniques:
• Conditions-of-use agreements
• Business key cross references
• Processing Log analysis
Tools:
• Data Integration Tools
• Data Remediation Tools
• Operational Data Stores
• Data Sharing Hubs
• Data Modeling Tools
• Metadata Repositories
• MDM Application Platforms
Metrics:
• Data Quality and Compliance
• Data Change Activity
• Data Ingestion and Consumption
• Data Sharing Availability
• Data Steward Coverage
• Data Sharing Volume and Usage
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 75 Context Diagram: Reference and Master Data

1.1 Business Drivers

For Master Data Management program:

- **Meeting organisational data requirements**: Multiple areas need to access same data sets, which confidence that they are complete, current and consistent
- **Managing data quality**: MDM enables a consistent representation of critical entities
- **Managing the costs of data integration:** Master Data reduces variation in identification and definition of critical data entities
- **Reducing risk:** Simplifies the data sharing environment

Centrally managed Reference Data enables the organisation to:

- Meet data requirements for multiple initiatives, reduce costs and risks of data integration
- Manage quality of Reference Data

1.2 Goals and Principles

Goals:

1. Ensuring the organization has complete, consistent, current, authoritative Master and Reference Data across organizational processes.
2. Enabling Master and Reference Data to be shared across enterprise functions and applications.
3. Lowering the cost and reducing the complexity of data usage and integration through standards, common data models, and integration patterns.

R and MDM guiding principles:

- **Shared data:** Managed to be sharable across organisation
- **Ownership**: Belong to the organisation. Require high level of stewardship
- **Quality**: Require ongoing monitoring and governance
- **Stewardship**: Business data stewards responsible
- **Controlled Change**:
    - **Master Data**: Represents the best view of currency and accuracy at any point in time. Caution when matching rules that change values. Should be reversible
    - **Reference Data**: Change follows defined process. Approve and communicate before implementing
- **Authority**: Master data values should be replicated only from the system of record

1.3 Essential concepts

1.3.1 Differences between Master and Reference Data

Malcolm Chisholm proposed a six-layer taxonomy of data (2008):

- Metadata
- Reference Data
- Enterprise Structure Data
- Transaction Structure Data
- Transaction Activity Data
- Transaction Audit Data

Chisholm's definition of Master Data (within the above taxonomy): An aggregation of Reference Data, enterprise structure data and transaction structure data.

- **Reference data:** Code and description tables used to categorise other data in the organisation. Relates data in the database to information outside the organisation
- **Enterprise Structure Data:** Business activity data e.g., chart of accounts
- **Transaction Structure data:** Describes things that must be present for a transaction to occur e.g., products, customers, vendors

DAMA Dictionary Definition (2009): Master Data is the data that provides the context for business activity data in the form of common and abstract concepts that relate to the activity. It includes the details (definition and identifiers) of internal and external objects involved in business transactions, such as customers, products, employees, vendors and controlled domains (code values).

David Loshin describes Master Data objects as core business objects used in different applications across the organisation along with their associated Metadata, attributes, definitions, roles, connections and taxonomies. Master Data objects represent those things that matter most to the organisation, that are logged in transaction, reported on, measured and analysed.

Master Data requires identifying and developing a trusted version of the truth for each instance of conceptual entities, and maintaining the currency of that version. Master Data Management works to resolve the differences in associations between data in different systems and processes to consistently identify individual entity instances in different contexts. This process must be managed over time so that the identifiers for these Master Data entity instances remain consistent.

- **Shared purposes of Reference and Master Data**:
  - Provide context to the creation and use of transactional data
  - Reference data provides context for Master Data
  - Enable data to be meaningfully understood
  - Shared resources managed at the enterprise level
- **Reference Data compared to Master Data sets:**
  - Less volatile
  - Fewer columns and rows than Transactional or Master Data sets
  - No entity resolution challenges

Focus of data management differs:

- **Master Data Management (MDM):** Control over Master Data values and identifiers that enable consistent use of the most accurate and timely data about essential business entities. Ensure availability of accurate current values while reducing risks of ambiguity
- **Reference Data Management (RDM):** Control over defined domain values and their definitions. Ensure the organisation has access to a complete set of accurate and current values for each concept represented

RDM is responsible for obtaining data and managing updates, as Reference Data can originate inside or outside the organisation. Ownership may be a challenge

1.3.2 Reference Data

Reference data is any data used to characterise or classify other data, or to relate data to information external to an organisation (Chisholm, 2001). Can be codes and descriptions or more complex hierarchies and mappings.

Common storage techniques:

- Code tables in relational databases linked by foreign keys

- Reference Data Management systems
- Object attribute specific Metadata to specify permissible values for APIs or user interface access

Reference Data Management entails control and maintenance of defined domain values, definitions and the relationships with and across domain values. The goal is to ensure values are consistent, current and accessible to the organisation.

1.3.2.1 Reference Data Structure:

Depends on the granularity and complexity.

1.3.2.1.1 Lists

Code value and a description which can be used in a drop down.

Code Value	Description
US	United States of America
GB	United Kingdom (Great Britain)

Table 17 Simple Reference List

Definitions added for a Help function.

Code	Description	Definition
1	New	Indicates a newly created ticket without an assigned resource
2	Assigned	Indicates a ticket that has a named resource assigned
3	Work In Progress	Indicates the assigned resource started working on the ticket
4	Resolved	Indicates request is assumed to be fulfilled per the assigned resource
5	Cancelled	Indicates request was canceled based on requester interaction
6	Pending	Indicates request cannot proceed without additional information
7	Fulfilled	Indicates request was fulfilled and verified by the requester

Table 18 Simple Reference List Expanded

1.3.2.1.2 Cross-Reference Lists

Used to translate code values of the same concept. May be at different granularities or same granularity with different values.

USPS State Code	ISO State Code	FIPS Numeric State Code	State Abbreviation	State Name	Formal State Name
CA	US-CA	06	Calif.	California	State of California
KY	US-KY	21	Ky.	Kentucky	Commonwealth of Kentucky
WI	US-WI	55	Wis.	Wisconsin	State of Wisconsin

Table 19 Cross-Reference List

ISO 3166-1 Alpha 2 Country Code	English Name	Local Name	Local Name Local Alphabet	French Name	...
CN	China	Zhong Guo	中国/中國	Chine	

Table 20 Multi-Language Reference List

1.3.2.1.3 Taxonomies

Taxonomic Reference Data structures capture information at different levels of specificity to support multifaceted navigation required by Business Intelligence.

Code Value	Description	Parent Code
10161600	Floral plants	10160000
10161601	Rose plants	10161600
10161602	Poinsettias plants	10161600
10161603	Orchid plants	10161600
10161700	Cut flowers	10160000
10161705	Cut roses	10161700

Table 21 UNSPSC (Universal Standard Products and Services Classification)

Code Value	Description	Parent Code
440000	Retail Trade	440000
445000	Food and Beverage Stores	440000
445200	Specialty Food Stores	445000
445210	Meat Markets	445200
445220	Fish and Seafood Markets	445200
445290	Other Specialty Food Stores	445200
445291	Baked Goods Stores	445290
445292	Confectionary and Nut Stores	445290

Table 22 NAICS (North America Industry Classification System)

1.3.2.1.4 Ontologies

Ontologies used to manage website content can be part of Reference Data, as they are used to characterise other data or relate organisational data to information beyond the boundaries of the organisation. Ontologies may also be considered Metadata.

1.3.2.2 Proprietary or Internal Reference Data

Reference data created within the organisation to support internal systems. RDM consists of managing them and ensuring consistency.

1.3.2.3 Industry Reference Data

Industry Reference Data describes data sets which are created and maintained by industry associations and government bodies in order to provide a standard for codifying important concepts.

1.3.2.4 Geographic or Geo-statistical Data

Enables classification or analysis based on geography.

1.3.2.5 Computational Reference Data

Used for common, consistent calculations. This data may change frequently, such as foreign exchange values.

1.3.2.6 Standard Reference Data Set Metadata

Maintain key Metadata about Reference Data sets to ensure their lineage and currency are understood and maintained.

Reference Data Set Key Information	Description
Formal Name	Official, especially if external name of the Reference Data set (e.g., ISO 3166-1991 Country Code List)
Internal Name	Name associated with the data set within the organization (e.g., Country Codes – ISO)
Data Provider	The party that provides and maintains the Reference Data set. This can be external (ISO), internal (a specific department), or external – extended (obtained from an external party but then extended and modified internally).
Data Provider Data Set Source	Description of where data provider's data sets can be obtained. This is likely a Universal Resource Identifier (URI) within or outside of the enterprise network.
Data Provider Latest Version Number	If available and maintained, this describes the latest version of the external data provider's data set where information may be added or deprecated from the version in the organization
Data Provider Latest Version Date	If available and maintained, this describes when the standard list was last updated
Internal Version Number	Version number of the current Reference Data set or version number of the last update that was applied against the data set
Internal Version Reconciliation Date	Date when data set was last updated based on the external source
Internal Version Last Update Date	Date data set was last changed. This does not mean reconciliation with an external version.

Table 23 Critical Reference Data Metadata Attributes

1.3.3 Master Data

Master Data is about **the key business entities** and should represent the **authoritative most accurate data available**, which can be trusted and used with confidence.

Business rules dictate the format and allowable values. Common organisational Master Data is data about:

- **Parties:** Individuals, organisations and their roles
- **Products and services:** Internal and external
- **Financial structures:** e.g., contracts, general ledger accounts
- **Locations:** e.g., addresses and GPS coordinates

1.3.3.1 System of Record, System of Reference

Where there are potentially different versions of the truth, we need to know more about the data to distinguish between them:

- **A System of Record** is an authoritative system where data is created/captured and maintained through a defined set of rules and expectations
- **A System of Reference** is an authoritative system where data consumers can obtain reliable data to support transactions and analysis. Examples are MDM applications, Data Sharing Hubs and Data Warehouses

1.3.3.2 Trusted Source, Golden Record

A **Trusted Source** is recognised as the "best version of the truth" based on a combination of automated rules and manual stewardship. (Any Master Data Management system)

A **Golden Record** represents 100% accurate data about an entity, also referred to as a "single version of the truth". Not always possible for multiple systems to have one version of the truth as there are usually multiple SOR. Promising data is Golden can be misleading, and can undermine confidence.

1.3.3.3 Master Data Management

Gartner's Definition: A technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared Master Data assets. Master Data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise, including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts.

MDM is a discipline made up of people, processes and technology.

Criteria to assess MDM requirements:

- Which roles, organisations, places and things are referenced repeatedly
- What data is used to describe people, places and things
- How the data is defined, structured, and the granularity
- Where data is created/sourced, stored, made available and accessed
- How it changes as it moves through systems
- Who uses the data, and for what purposes
- Criteria used to understand the quality and reliability of the data and its sources

Planning for Master Data Management within a domain:

- Identify candidate sources that will provide a comprehensive view of Master Data entities
- Develop rules for accurately matching and merging entity instances
- Establish an approach to identify and restore inappropriately matched and merged data
- Establish an approach to distribute trusted data systems across the enterprise

1.3.3.3.1 Master Data Management Key Processing Steps

Figure 76 Key Processing Steps for MDM

- **Data Model Management:** Clear and consistent definitions make sense to business at the enterprise level
- **Data Acquisition:** Data representing the same entity can look different. Plan for acquiring new data as a reliable repeatable process. High level cleansing tools, and matching rules, then perform data quality on the new data

Source ID	Name	Address	Telephone
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, DA	2345678900
345	Jane Smith	123 Main, Dataland, DA	234-567-8900

Table 24 Source Data as Received by the MDM System

- **Data Validation, Standardisation and Enrichment:** Reduce variation in format and reconcile values:
  - **Validation:** Identify clearly incorrect or defaulted data
  - **Standardisation:** Data conforms to standard Reference Data values and formats
  - **Enrichment:** Add attributes that improve entity resolution services

Source ID	Name	Address (Cleansed)	Telephone (Cleansed)
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900

Table 25 Standardised and enriched Input Data

- **Entity Resolution and Identifier Management:**

  Entity resolution is the process of determining whether two references to real world objects refer to the same object or different objects.

  Activities:

  - **Matching:** or candidate identification is the process of identifying how different records relate to a single entity. Use similarity analysis to avoid false positives or negatives
    - **Deterministic Algorithms:** Parsing and standardisation rely on patterns
    - **Probabilistic Algorithms:** Rely on statistical techniques
  - **Identity Resolution:** Keep a history of matches so that those less confident matches due to conflicting values can be undone if found to be incorrect

Source ID	Name	Address (Cleansed)	Telephone (Cleansed)	Candidate ID	Party ID
123	John Smith	123 Main, Dataland, SQ 98765		XYZ	1
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	XYZ, ABC	2
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	ABC	2

Table 26 Candidate Identification and Identity Resolution

  - **Matching Workflows / Reconciliation Types:** Different scenarios require different workflows:
    - **Duplicate identification match rules:** Specific set of data elements that uniquely identify an entity and identify merge opportunities
    - **Match-link rules:** Identify and cross-reference records that appear to relate to the master record without updating the content of the cross-referenced record
    - **Match-merge rules:** Match records and merge data into a single unified and comprehensive record. Complex

- **Master Data ID Management:** Two types of identifiers managed in an MDM environment:
  - **Global ID** is the MDM solution assigned unique identifier attached to reconciled records. Should be automatically generated
  - **X-Ref Management** is management of the relationship between source IDs and the Global ID
- **Affiliation Management:** Establishing and maintaining relationships between Master Data records of entities that have real-world relationships. Data Architecture design of the MDM system which kind of relationship to leverage between entities:
  - **Affiliation relationships** are programmed and are the most flexible
  - **Parent-child relationships** have implied hierarchical navigation structure

- **Data Sharing and Stewardship:** Data Stewards resolve incorrectly matched situations and improve matching algorithms

1.3.3.4 Party Master Data

Data about individuals, organisations and the role they play in business relationships. Examples from different environments:

- **Commercial:** Customers, employees, vendors, partners, competitors
- **Public sector:** Citizens
- **Law enforcement:** Suspects, witnesses, victims
- **not for profit:** Members, donors
- **healthcare:** Patients, providers
- **Education:** Students, faculty

Customer Relationship Management (CRM) systems manage Master Data about customers.

Master Data is challenging for parties playing more than one role in an organisation.

1.3.3.5 Financial Master Data

Data about business units, cost centres, profit centres, general ledger accounts, budgets, projections and projects. The central hub of financial Master Data is an Enterprise Resource Planning (ERP) system.

1.3.3.6 Legal Master Data

Data about contracts, regulations and other legal matters.

1.3.3.7 Product Master Data

Can focus on the organisation's products and services or on industry-wide products and services. Different types of product Master Data solutions:

- **Product Lifecycle Management (PLM):** Managing the lifecycle of a product/service from conception to disposal
- **Product Data Management (PDM):** Engineering and manufacturing. Enables secure sharing of product information such as design drawings (CAD)
- **Product Data in Enterprise Resource Planning (ERP):** SKUs to support order entry to inventory
- **Product Data in Manufacturing Execution Systems (MES):** Raw inventory to finished goods
- **Product Data in Customer Relationship Management (CRM):** Marketing, sales and support interactions

1.3.3.8 Location Master Data

The ability to track and share geographic information and to create hierarchical relationships based on geographic information.

- **Location Reference Data** is usually geopolitical data handled by external organisations
- **Location Master Data** are address related to parties and businesses

1.3.3.9 Industry Master Data – Reference Dictionaries

Authoritative listings of Master Data entries that can be purchased. They can be licensed and can provide a starting point for matching and linking new records.

1.3.4 Data Sharing Architecture

Each Master Data subject area usually has its own system of record e.g., CRM or ERP systems. Hub-and-spoke model for sharing Master Data:

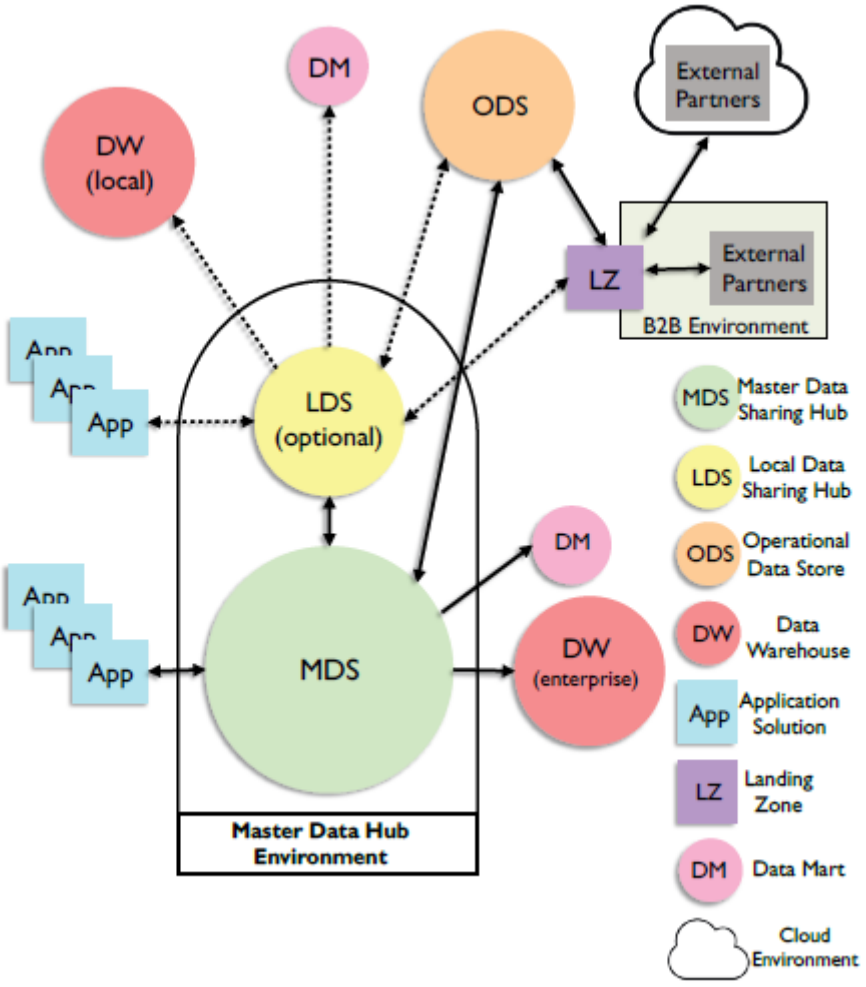


Figure 77 Master Data Sharing Architecture Example

Three approaches to implementing Master Data hub environment:

- **Registry:**
  - An index that points to Master Data in various systems of record
  - Manage Master Data local to their applications
  - Easy to implement, but complex queries are challenging
  - Multiple business rules need to be implemented

- **Transaction Hub:**
  - Applications interface with the hub to access and update Master Data which exists **only** within the Transaction Hub and not in the applications
  - The Hub is the **system of record** for Master Data
  - Better governance and consistency, but costly
  - Business rules are implemented in the Hub
- **Consolidated approach**:
  - Systems of record manage Master Data for their applications
  - It is also consolidated within a common repository (replication)
  - There is no need to access directly from the systems of record
  - There will be latency between the Hub and SOR

2 Activities

2.1 MDM Activities

- **Define MDM Drivers and Requirements:**
  - Easier to define requirements for an application than the whole enterprise
  - Prioritise Master Data efforts on cost benefit of the proposed improvements
  - Start with simplest category to learn from the process
- **Evaluate and Assess Data Sources**
  - Understand the structure of existing data and how it is collected or created
  - Understand the quality of the data, as poor quality data complicates a Master Data project
  - Assess disparity between sources
  - May be able to purchase standardised data such as Reference Directories
- **Define Architectural Approach:** Depends on business strategy, the platforms for existing sources and the lineage and volatility of the data
- **Model Master Data:** As Master Data is an integration process, model data within subject areas. A logical or canonical model
- **Define Stewardship and Maintenance Processes:** Technical solutions still require the oversight of Data Stewards to address records that fall out of the process and why
- **Establish Governance Policies to enforce use of Master Data:** The benefits come once people start using the Master Data

2.2 Reference Data Activities

- **Define Drivers and Requirements:**
  - **Drivers:** Operational efficiency and higher data quality
  - **Requirements:** Driven by the most important Reference Data sets
- **Evaluate and Assess Data Sources:**
  - **External:** Vendor that guarantees updates on a schedule and ensures quality data
  - **Internal:** Owners should understand the benefits of central management of their data sets
- **Define Architectural Approach:** Tool should allow for manual updates
- **Model Reference Data sets:** Models are Metadata which helps consumers understand the relationships within the Reference Data sets, and the data quality rules
- **Define Stewardship and Maintenance Processes:** Capture Metadata about reference data sets

- **Establish Reference Data Governance Policies:** Govern quality and mandate use of RD

3 Tools and Techniques

MDM requires identity management enabled tools:

- Data integration tools
- Data remediation tools
- Operational data stores (ODS)
- Data sharing hubs (DSH)
- Specialised MDM applications

4 Implementation Guidelines

As Master and Reference Data Management are forms of data integration, the same implementation principles that apply to Data Integration and Interoperability.

Implement incrementally through a series of projects defined in an implementation roadmap, prioritised on business needs and guided by an overall architecture.

It is vital to have data governance professionals who understand the challenges of RDM and MDM and can assess the maturity and ability of the organisation to meet them.

4.1 Adhere to Master Data Architecture

The integration process should take into account:

- The organisational structure of the business
- The number of distinct systems of record
- The data governance implementation
- The importance of access and latency of data values
- The number of consuming systems and applications

4.2 Monitor Data Movement

Monitor data as it flows within the Reference or Master Data sharing environment to:

- Show how data is used across the organisation
- Identify data lineage from/to administrative systems and applications
- Assist root cause analysis of issues
- Show effectiveness of data ingestion and consumption integration techniques
- Denote latency of data values from source systems through consumption
- Determine validity of business rules and transformations executed within integration components

4.3 Manage Reference Data Change

Reference Data is a shared resource; therefore, it should not be locally controlled, but channels to receive and respond to change requests must be provided according to policies and procedures put in place by the Governance Council.

Planned/scheduled changes such as periodic updates to industry codes require less governance than ad hoc changes.

Figure 78 Reference Data Change Request Process

4.4 Data Sharing Agreements

Data sharing agreements stipulate what data can be shared and under what conditions. Helps when issues regarding quality of data brought in or availability arise. Driven by the Data Governance function and involves Data Architects, Data Providers, Data Stewards, Application Developers, Business Analysts, Compliance/Privacy Officers and Security Officers.

SLAs should be in place so that the quality data can be provided to downstream consumers.

5 Organisation and Cultural Change

It is not easy for people to relinquish control of their data to create shared resources. People may perceive MDM and RDM efforts as adding complications to their processes.

The most challenging cultural change is determining which individuals are accountable for which decisions.

6 Reference and Master Data Governance

Because they are shared resources, Reference and Master Data require governance and stewardship. Governance processes will determine:

- The data sources to be integrated
- Data quality rules to be enforced
- Conditions of use rules
- Activities to be monitored and the frequency of monitoring
- Priority and response levels of stewardship efforts
- How information is represented to meet stakeholder needs
- Standard approval gates, expectations in RDM and MDM deployment

Governance processes bring compliance and legal stakeholders together with information consumers to ensure risks are mitigated through definition and incorporation of privacy, security and retention policies.

Data Governance must have the ability to review, receive and consider new requirements and changes. Make principles, rules and guidelines available to all using Reference and Master Data.

6.1 Metrics

- **Data quality and compliance:** DQ dashboards
- **Data change activity:**
  - Metrics denote rate of change of data values
  - Provide insight to systems providing data to sharing environment
  - Used to tune algorithms in MDM process

- **Data ingestion and consumption:** Denote and track data contributing systems and what business areas are subscribing to shared data
- **Service Level Agreements:** Level of adherence to SLAs provides insight to data and technical problems
- **Data Steward coverage:** Used to identify gaps in support
- **Total cost of ownership:** Must be consistently applied across the organisation to be effective
- **Data sharing volume and usage:** The volume and velocity of data defined, ingested and subscribed to and from the data sharing environment

Chapter 11

Data Warehousing & Business Intelligence

The **Data Warehouse** is meant to enable decision support systems to share core enterprise data from a common data model. The **Enterprise Data Warehouse** reduces data redundancy, improves the consistency on information, and enables the enterprise to make better decisions.

1 Introduction

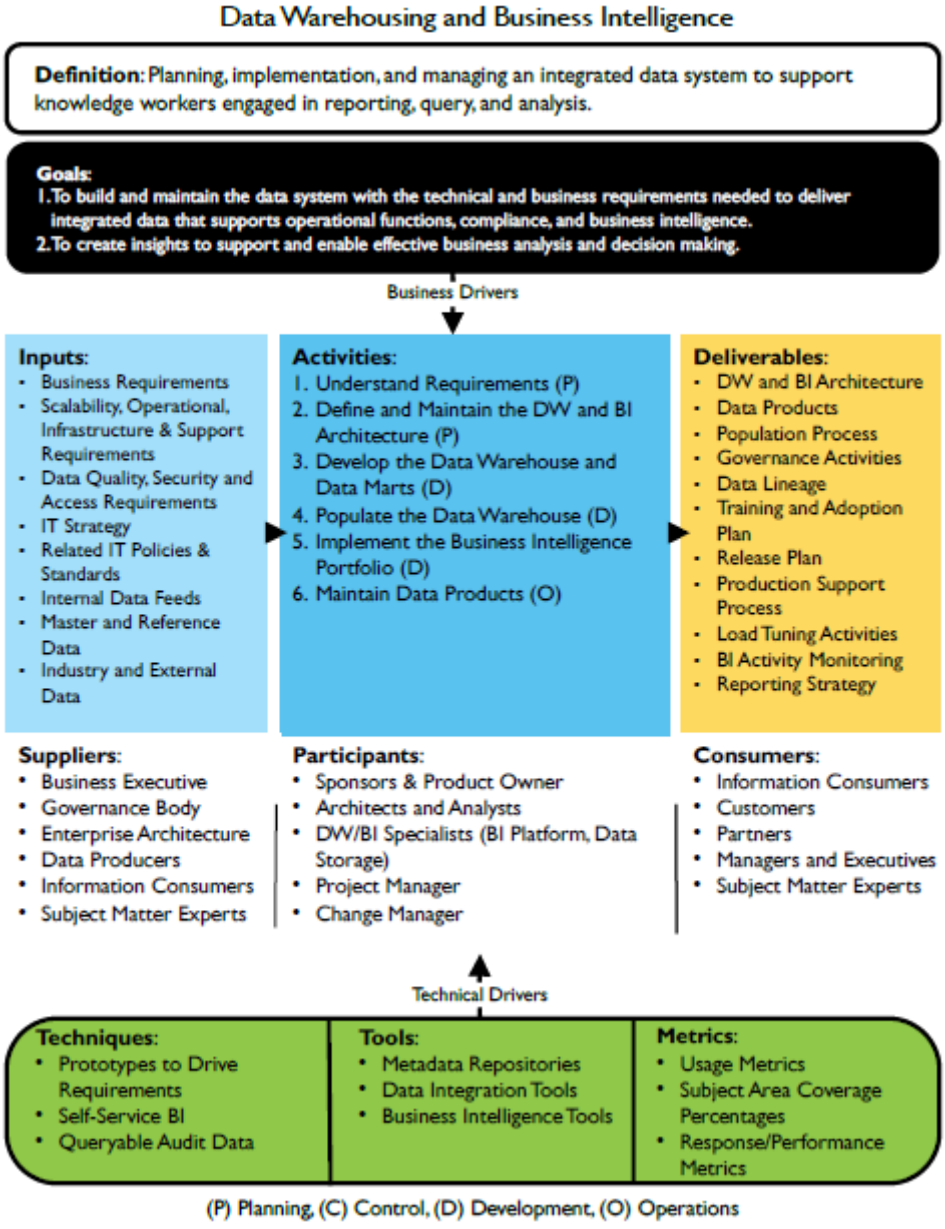


Figure 79 Context Diagram: DW/BI

1.1 Business Drivers

- Business Intelligence support
- Compliance requirements
- Support operational activities

- Enables effective business analysis and decision-making
- Find ways to innovate based on insights from data

1.2 Goals and Principles

Goals:
1. To build and maintain the data system with the technical and business requirements needed to deliver integrated data that supports operational functions, compliance, and business intelligence.
2. To create insights to support and enable effective business analysis and decision making.

Organisations implement data warehouses to:

- Support BI activity
- Enable effective business analysis and decision-making
- Find ways to innovate based on insights from data

Principles to implement a Data Warehouse:

- **Focus on business goals:** DW serves organisational priorities and solves business problems
- **Start with the end in mind:** Business priority and scope end-data-delivery in BI space drives creation of DW content
- **Think and design globally; act and build locally:** Architecture guided by end-vision, but build and deliver in sprints to enable return on investment
- **Summarise and optimise last, not first:** Build on atomic data
- **Promote transparency and self-service:** Provide more context (Metadata) and keep users informed of updates and changes
- **Build Metadata with the warehouse:** To be able to explain the data, capture as part of the development cycle and manage as an ongoing activity
- **Collaborate:** With other data initiatives especially DG, DQ and Metadata
- **One size does not fit all:** Different groups of data consumers need different tools

1.3 Essential Concepts

1.3.1 Business Intelligence

Two meanings:

- **Type of data analysis** aimed at understanding organisational activities and opportunities
- **A set of technologies** that enable decision support analysis (querying, data mining, statistical analysis, reporting, scenario modelling, data visualisation and dashboarding)

1.3.2 Data Warehouse

A Data Warehouse (DW) is a combination of two components:

- Integrated decision support database
- Software programs used to collect, cleanse, transform and store data from a variety of internal and external sources

An **Enterprise Data Warehouse** (EDW) is a centralised data warehouse designed to service the BI needs of the entire organisation. Adheres to the Enterprise Data Model.

1.3.3 Data Warehousing

Data warehousing describes the operational extract, cleansing, transformation, control and load processes that maintain the data in a data warehouse. Integrated, historical, enforces business rules and maintains business relationships. Processes interact with Metadata repositories.

- **Structured data:** elements in defined fields as documented in data models
- **Semi-structured data:** Electronic elements organised as semantic entities with no required attribute affinity
- **Unstructured data:** Not predefined through a data model

1.3.4 Approaches to Data Warehousing

Two thought leaders – **Bill Inmon** and **Ralph Kimball**:

- **Inmon**:
  - A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making processes
  - Normalised, relational model
- **Kimball**:
  - A copy of transaction data specifically structured for query and analysis
  - Dimensional model

Core ideas:

- Warehouses store data from other systems
- Storage includes organising data in ways that increases its value
- Warehouses make data accessible and useable for analysis
- Organisations build warehouses to make reliable, integrated data available to authorised stakeholders
- Warehouse data serves many purposes

1.3.5 Corporate Information Factory (Inmon)

CIF Illustrates the differences between warehouses and operational systems

- **Subject oriented:** Based on major business entities, not function or application
- **Integrated:** The warehouse becomes a system of record for the data
- **Time variant:** Stores data as it exists at a set point in time
- **Non-volatile:** New records are appended not updated
- **Aggregate and detail data:** Details of atomic level transactions as well as summarised data
- **Historical:** The focus of operational systems is current data. Warehouses contain lots of historical data as well

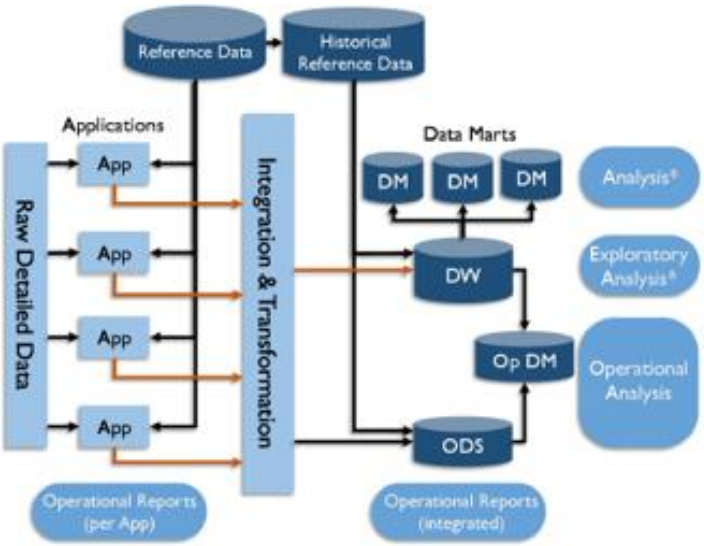


Figure 80 The Corporate Information Factory

CIF Components:

- **Applications:** Applications perform operational processes and bring data into the DW and operational data store (ODS) where it can be analysed
- **Staging area:** A database between the operational and target where the ETL takes place. Data is transient.
- **Integration and transformation:** Data from disparate sources is transformed in the integration layer into the standard corporate representation model in the DW and ODS
- **Operational Data Store (ODS):** An integrated relational database of operational data, containing current or near-term data, volatile. Meet the need for low latency data. Can be the primary source for the data warehouse
- **Data marts:** Provide data prepared for analysis. Sub-set of the warehouse designed for the needs of specific consumers. Data marts are designed using dimensional modelling and denormalisation
- **Operational Data Mart (OpDM):** Sourced from ODS, volatile and focused on tactical decision support
- **Data Warehouse:** A single integration point for corporate data to support management decision making and strategic analysis and planning
- **Operational reports:** The output from the data stores
- **Reference, Master and external data:** Data required to understand data from applications, and simplifies integration to DW

Comparison between data stored in DW and Marts and Data on operational systems:

Data in DW and Marts	Data in Operational Systems
• Organised by subject	• Organised by function
• Integrated	• Siloed
• Data is time-variant	• Current value only
• High latency	• Lower latency
• Significantly more historical data	• Less historical data

1.3.6 Dimensional DW (Kimball)

Kimball: "a copy of transaction data specifically structured for query and analysis".

Warehouse data is stored in a **dimensional model**, not normalised, enabling consumers to use the data and optimise query performance.

Star Schema or **dimensional** models:

- **Facts** contain quantitative data about business processes. (The fact table is also called a meter which contains measures – Hoberman)
- **Dimensions** store descriptive attributes related to fact data and allow consumers to answer questions about the facts

One Fact table is joined to many dimensions – looks like a star.

Conformed dimensions are common dimensions and are shared by multiple fact tables via a bus. Multiple data marts can also be integrated at enterprise level by plugging into the bus of conformed dimensions.

	Subject Areas				
Business Processes	**Date**	**Product**	**Store**	**Vendor**	**Warehouse**
Sales	X	X	X		
Inventory	X	X	X	X	X
Orders	X	X		X	
Conformed Dimension Candidate	*Yes*	*Yes*	*Yes*	*Yes*	*No*

Table 27 DW-Bus Matrix Example

Kimbal's Data Warehouse Chess Pieces is more expansive than Inmon's:

- **Operational source systems**: Equivalent to the applications in CIF
- **Data Staging area:** Processes needed to integrate and transform data for presentation. Kimball's EDW
- **Data presentation area:** Like data marts in the CIF. The DW Bus conformed dimensions
- **Data access tools:** End users' needs drive adoption of data access tools

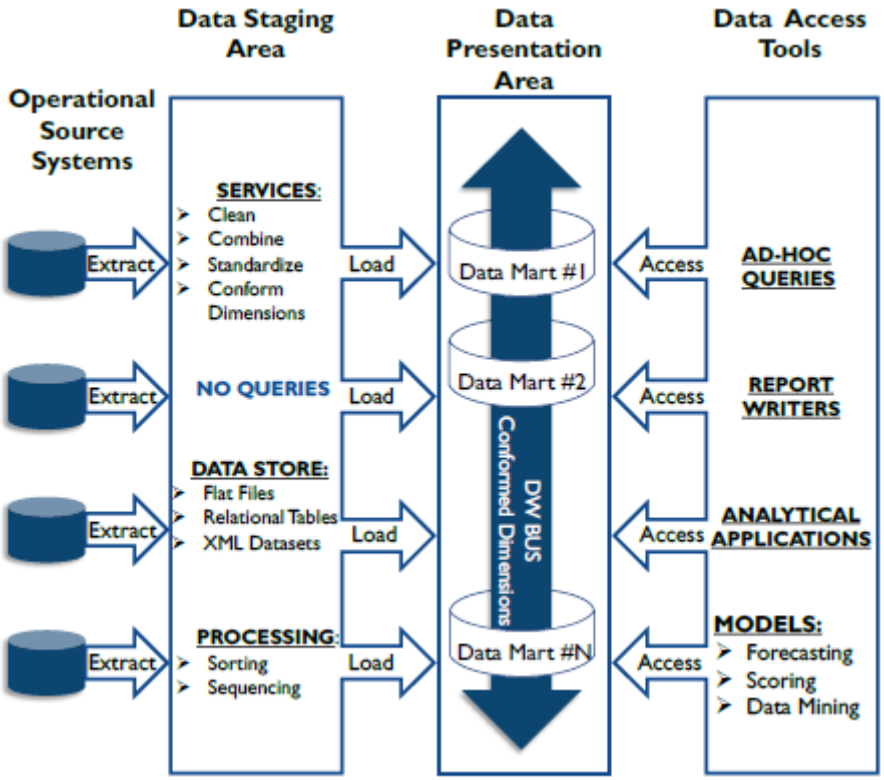


Figure 81 Kimball's Data Warehouse Chess Pieces

1.3.7 DW Architecture Components

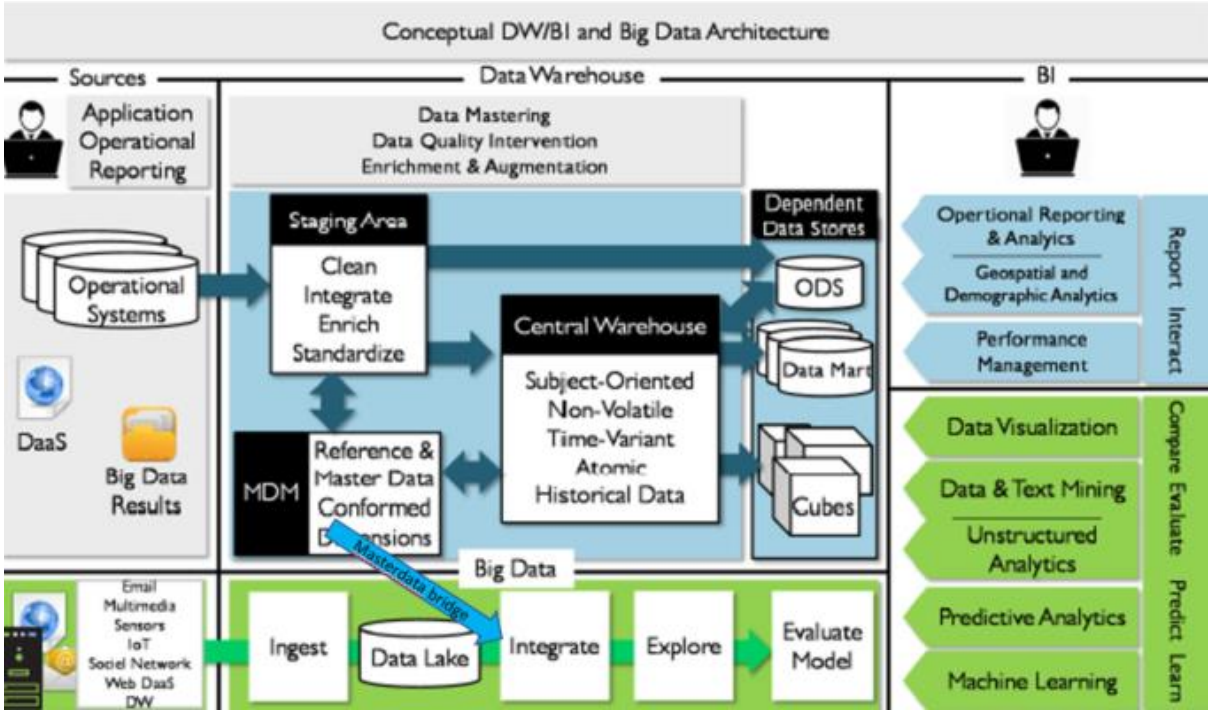


Figure 82 Conceptual DW/BI and Big Data Architecture

The evolution of Big Data has added to the DW/BI landscape by providing another data entry path. Lifecycle is depicted:

- **Source Systems:** Operational systems and external data to be brought into the DW/BI environment
- **Data Integration:** ETL, data virtualisation and other techniques to get data into a common form and location. (Arrows represent the integration process)
- **Data Storage Areas:** The warehouse has a set of storage areas
    - **Staging area:** Intermediate storage for ETL and integration
    - **Reference and Master Data conformed dimensions:** May be separate repositories
    - **Central warehouse:** DW Data usually persists in the central or atomic layer, which maintains all historic atomic data as well as the latest batch run. Data structure based on performance needs and use patterns:
        - Relationship between Business and surrogate keys for performance
        - Indexes and foreign keys to support dimensions
        - Change Data Capture (CDC) techniques that detect, maintain and store history
    - **Operational Data Store (ODS):** Version of the central persisted store that supports lower latencies and therefore operational use
    - **Data marts:** Type of data store often used to support presentation layers of the data warehouse environment. Can also be a functional subset of the DW
    - **Cubes:** Three classic implementation approaches support Online Analytical Processing (OLAP). Their names relate to underlying database types, such as Relational (ROLAP), Multi-dimensional (MOLAP) and Hybrid (HOLAP)

1.3.8 Types of Load Processing

Historical (loaded once) and ongoing updates (consistently scheduled)

- **Historical Data:** Data warehouse captures detailed history of the data it stores
    - **Inmon:** All data stored in single DW layer with common integration and transformation layer. (Need Enterprise Model for success)
    - **Kimball:** DW is a combination of departmental data marts which store history at atomic level
    - **Data Vault:** Hybrid between 3NF and Star Schema (See DM & D Chapter 5). Hubs (PK), Links and Satellites. Facts persist as atomic structures
- **Batch Change Data Capture:** Different change capture techniques:

Method	Source System Requirement	Complexity	Fact Load	Dimension Load	Overlap	Deletes
Time stamped Delta Load	Changes in the source system are stamped with the system date and time.	Low	Fast	Fast	Yes	No
Log Table Delta Load	Source system changes are captured and stored in log tables	Medium	Nominal	Nominal	Yes	Yes
Database Transaction Log	Database captures changes in the transaction log	High	Nominal	Nominal	No	Yes
Message Delta	Source system changes are published as [near] real-time messages	Extreme	Slow	Slow	No	Yes
Full Load	No change indicator, tables extracted in full and compared to identify change	Simple	Slow	Nominal	Yes	Yes

Table 28 CDC Technique Comparison

- **Near-real-time and Real-time:** Operational BI requires lower latency and the inclusion of volatile data in the warehouse
    - **Trickle feeds (source accumulation):** Batch loads on a more frequent schedule, or when a threshold is reached
    - **Messaging (Bus accumulation):** Small messages are published to a bus when they occur. Used by Data-as-a-Service (DaaS)
    - **Streaming (Target accumulation):** A target system collects data as it is received into a buffer or queue

2 Activities

2.1 Understand Requirements

Operational systems depend on precise, specific requirements. A data warehouse brings together data that will be used in many systems to explore and analyse.

2.2 Define and Maintain the DW/BI Architecture

DW/BI architecture describes where the data comes from, where it goes, when it goes, why and how it goes into the warehouse. The technical requirements include performance, availability and timing needs.

2.2.1 Define DW/BI Technical Architecture

DW/BI architectures should have a mechanism to connect back to the transactional and operational reports in an atomic DW.

A **conceptual model** aligns with business needs. Test by prototyping. Validate with the Enterprise Data Model.

2.2.2 Define DW/BI Management Processes

Address production management with a coordinated and integrated maintenance process, delivering regular releases to the business community. Establish a release schedule to manage each update to the deployed data product as a software release.

2.3 Develop the Data Warehouse and Data Marts

Three concurrent development tracks:

- **Data:** The data business requires for the analysis it wants to do:
  - Identify best sources
  - Remediation, transformation, integration, storage and availability rules
  - How to handle data that does not fit
- **Technology**: Back end systems and processes supporting data storage and movement
- **Business Intelligence tools**

2.3.1 Map Sources to Targets

Source to target mapping establishes transformation rules for data elements from individual sources to the target system. A solid taxonomy is necessary. It is often the **logical model**.

2.3.2 Remediate and Transform Data

Cleansing activities enforce standards to correct and enhance the domain values of individual data elements. Should be done in source systems.

2.4 Populate the Data Warehouse

The DW/BI team must publish clear rules for what data detail the DW contains, and what will be available via only operational reporting.

Key factors to consider when defining the population approach:

- Required latency
- Availability of sources
- Batch windows or upload intervals
- Target databases
- Dimensional aspects
- Timeframe consistency of the data warehouse and data mart
- Data quality processing
- Time to perform transformations, and late-arriving dimensions and data rejects
- Change data capture processes

2.5 Implement the Business Intelligence Portfolio

Identify the right tools for business communities – find similarities through alignment of common business processes, analyses, management styles and requirements.

- **Group users according to needs:** Know the user groups, then match the tool.IT Developers extracting data. Information consumers. Users' needs may change according to roles and skills
- **Match tools to user requirements**: Imbedded analytics. Virtualisation. BI Suites

2.6 Maintain Data Products

An implemented data warehouse and its customer-facing BI tools is a data product. Enhancements and extensions should be implemented incrementally.

- **Release Management:** Release management is critical to an incremental development process that grow new capabilities. This process keeps the warehouse up-to-date, clean and performing at its best. Below a quarterly schedule is illustrated

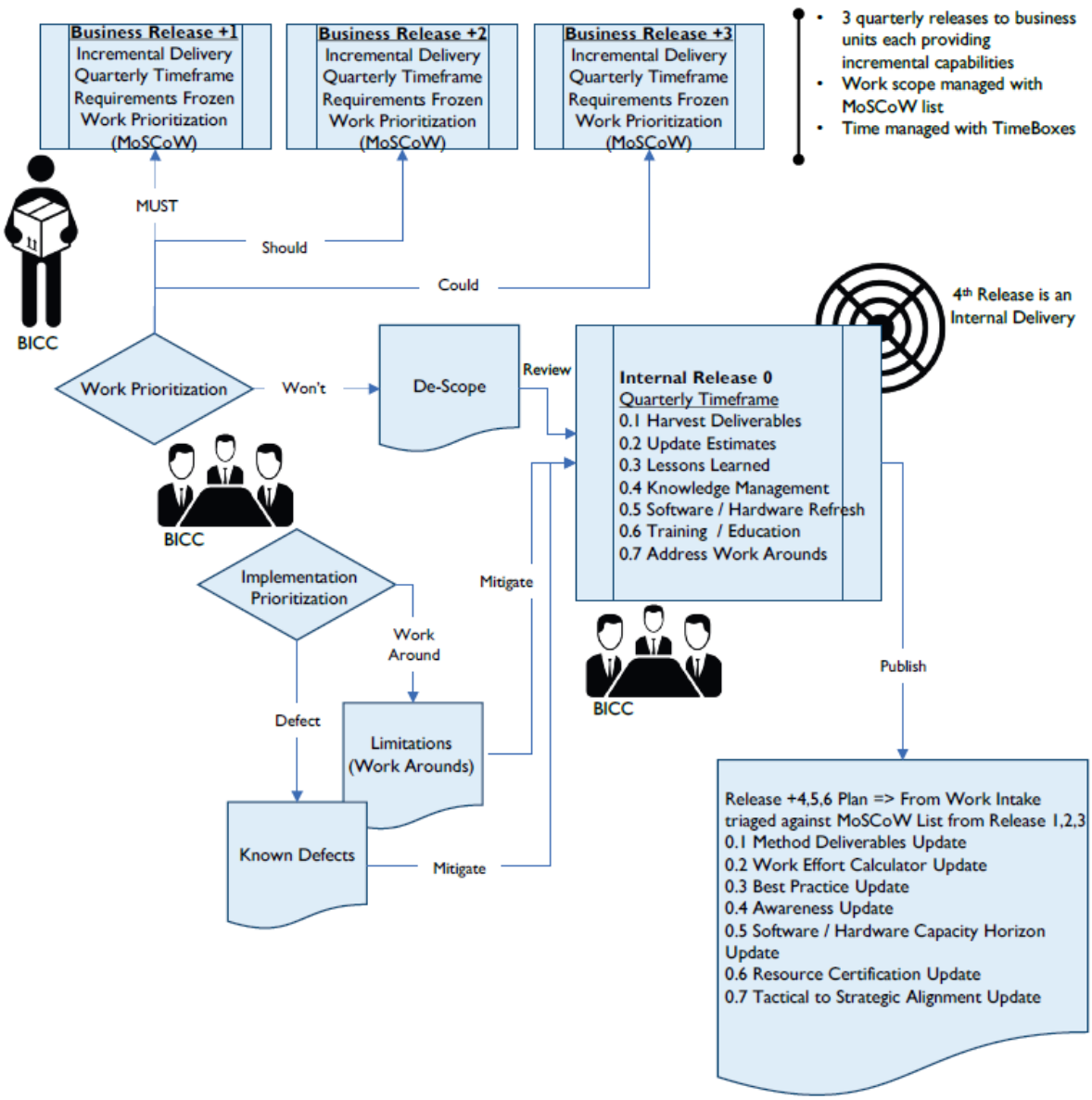


Figure 83 Release Process Example

- **Manage Data Product Development Lifecycle:** While the consumers are using the DW, the DW team is preparing the next iteration, which can extend an existing increment or add new functionality by onboarding a new business unit
- **Monitor and tune load processes:** Monitor for bottlenecks and dependencies. Tune the database. Users consider the warehouse an active archive
- **Monitor and tune BI activity and performance:** Customer-facing satisfaction metrics

3 Tools

3.1 Metadata Repositories

Automate and integrate population

3.1.1 Data Dictionary / Glossary

- Necessary for use of DW
- Describes the data in business terms and other information needed to use data

- Comes from the logical model

3.1.2 Data and Data Model Lineage

Integration tools offer lineage analysis for the population code and the physical model and database. Uses for documented data lineage:

- Investigation of root causes of issues
- Impact analysis of changes
- Determine reliability of data based on origin

3.2 Data Integration Tools

Used to populate the warehouse. Also schedule jobs that account for complex data delivery from many sources

3.3 Business Intelligence Tools Types:

- **Operational Reporting:** Analyse business trends to discover patterns. **Tactical BI** to support short term decisions
- **Business Performance Management:** Formal assessment of metrics aligned with organisational goals. **Strategic BI** to support long term corporate goals and objectives
- **Descriptive, self-service analytics:** BI to the **front line of business**. Service-oriented Architecture (SOA) and Big Data
- **Operational Analytic Applications:** Differ from OLAP and BI tools as they include processes to extract data from source systems, a data mart and pre-built reports and dashboards
- **Multi-dimensional Analysis – OLAP:** Online Analytical processing
    - **Slice:** A subset of a multidimensional array corresponding to a single value for one or more members not in the subset
    - **Dice:** A slice on more than two dimensions of a data cube or more than two consecutive slices
    - **Drill down/up:** User navigates among levels of data from most summarised (up) to most detailed (down)
    - **Roll-up:** Compute all the data relationships for one or more dimensions
    - **Pivot:** Changes the dimensional orientation of a report or page display
    - **3 implementation processes:**
        - **Relational Online Analytical Processing (ROLAP):** Supports OLAP using multidimensionality in the two-dimensional tables of RDBMS
        - **Multi-dimensional Online Analytical Processing (MOLAP):** Supports OLAP by using specialised multi-dimensional database technology
        - **Hybrid Online Analytical Processing (HOLAP):** Combination of ROLAP and MOLAP

4 Techniques

- Prototypes to drive requirements
    - Profile the data:
        - Reduce risk of unexpected data
        - Disclose differences in sources that may be obstacles in integration
- Self-service BI
- Queryable data for audit

5 Implementation Guidelines

- **Readiness assessment / Risk Assessment:** May be a gap between embracing venture and ability to sustain it
    - Pre-requisite checklist:
        - Has business support
        - Aligned with strategy
        - Has defined architectural approach
    - Define data sensitivity and security constraints
    - Select tools
    - Secure resources
    - Create source data ingestion process
    - Identify sensitive or restricted data that may need masking
    - Ensure data governance processes for review and approval have been followed
- **Release Roadmap:** Suggested approach:
    - Incremental leveraging of DW Bus matrix as a marketing and communication tool
    - Use business-determined priorities tethered to exposure metrics to determine how much rigor and exposure to apply to each increment
    - Apply consistent needs and abilities processes to determine next business unit to onboard
    - Maintain a back-order work item list to identify outstanding capabilities
    - Determine technical difficulties which may alter the order of delivery
    - Package work into a software release
    - Agree on pace of release (quarterly, monthly, weekly or faster if appropriate)
    - Roadmap of release dates to manage with business partners
- **Configuration Management:**
    - Aligns with release roadmap
    - Scripts to automate development, testing and transport to production
    - Provides version control as it brands the release at the database level
    - Automated and manually generated programs are tied to that brand
- **Organisation and Cultural Change:** Keep business focus throughout the DW/BI lifecycle. Align projects behind real business needs and assess necessary business support. Critical success factors are:
    - **Business sponsorship:** DW/BI projects require strong executive sponsorship, an engaged steering committee and commensurate funding
    - **Business goals and scope:** Clearly identified business need
    - **Business resources:** Management commitment to resource
    - **Business readiness:**
        - Business prepared for long term incremental delivery
        - Business committed to establishing centres of excellence to sustain product
        - Breadth of knowledge or skill gap within target community that can be crossed within a single increment
    - **Vision alignment:** How well does the IT strategy support business vision?
    - **Dedicated Team:** To manage ongoing operations of the production environment:
        - **Front office group** to notify maintenance team of deficiencies to be addressed
        - **Back office** support team ensures production configuration has executed as required

6 DW/BI Governance

Governance activities should be completed and addressed during implementation. Specific governance deliverables can be added to the Software Development Lifecycle. Warehouse governance processes should be:

- Aligned with risk management
- Business driven as different business units have different needs
- Mitigate risks (not curtail execution)

Most critical functions:

- Those that govern the business operated discovery or refinement area
  - Handshaking
  - Instantiate data
  - Transfer data
  - Discard data
  - Data archival and time horizons for boundaries to prevent sprawl
- Those that ensure quality within the warehouse
  - Assign time, resources and programs to remediate data
- One-off events: part of lifecycle but curtail them in the pilot area
- Policies required for procedures in real-time environment
- Risk exposure mitigation matrix. Curtail risk with governance functions

6.1 Enabling Business Acceptance

Sign-off by business is part of User Acceptance Testing, and is paramount for every DW/BI implementation. Critically important architectural sub-components:

- **Conceptual Data Model:** Key business concepts and how they are related to each other
- **Data quality feedback loop:**
  - How data issues are identified and remediated
  - Owners of the systems where they originated informed and held accountable for fixing them
  - How are issues caused by the DW data integration process remediated?
- **End-to-end Metadata:** How does the architecture support integrated end-to-end flow of Metadata? Is access to meaning and context part of the architecture?
- **End-to-end verifiable data lineage:** Is a system of record identified for all data?

6.2 Customer / User Satisfaction

Collecting, understanding and acting on customer feedback can be facilitated through regular meetings with user representatives.

6.3 Service Level Agreements

Specify business and technical requirements.

6.4 Reporting Strategy

Reporting strategy across the BI portfolio must address:

- Standards, processes, guidelines, best practices and procedures
- Security access to sensitive data elements to only entitled users

- Access mechanisms to describe how users want to interact, report, examine or view data
- User community type and appropriate tool to consume it with
- Nature of reports (summary, detailed, exception) and frequency, timing, distribution, storage formats
- Use of visualisation capabilities
- Trade-offs between timeliness and performance

Evaluate regularly to see if they are still providing value.

Data Source governance, monitoring and control are vital.

Centre of Excellence can empower business users to the self-service model.

6.5 Metrics

- **Usage Metrics:**
  - Number of registered, connected or concurrent users
  - Number of user accounts for each tool
  - Number of queries per user community per timeframe
- **Subject area coverage percentages**
  - How much of the warehouse is being accessed by each department
  - Mapping operational sources to targets
- **Response and performance metrics**
  - Response times can be retrieved from query tools
  - Load times for each data product in raw format
  - Query records, data refresh and data extract times for objects provided to users
  - Use the metrics to validate or adjust service levels

Chapter 12
Metadata Management

1 Introduction

Metadata helps an organisation understand its data, systems and workflows. Metadata includes information about:

- Technical and business processes
- Data rules and constraints
- Logical and physical data structures
- Describes the data (databases, data elements, data models)
- Concepts the data represents
- Relationships between data and concepts

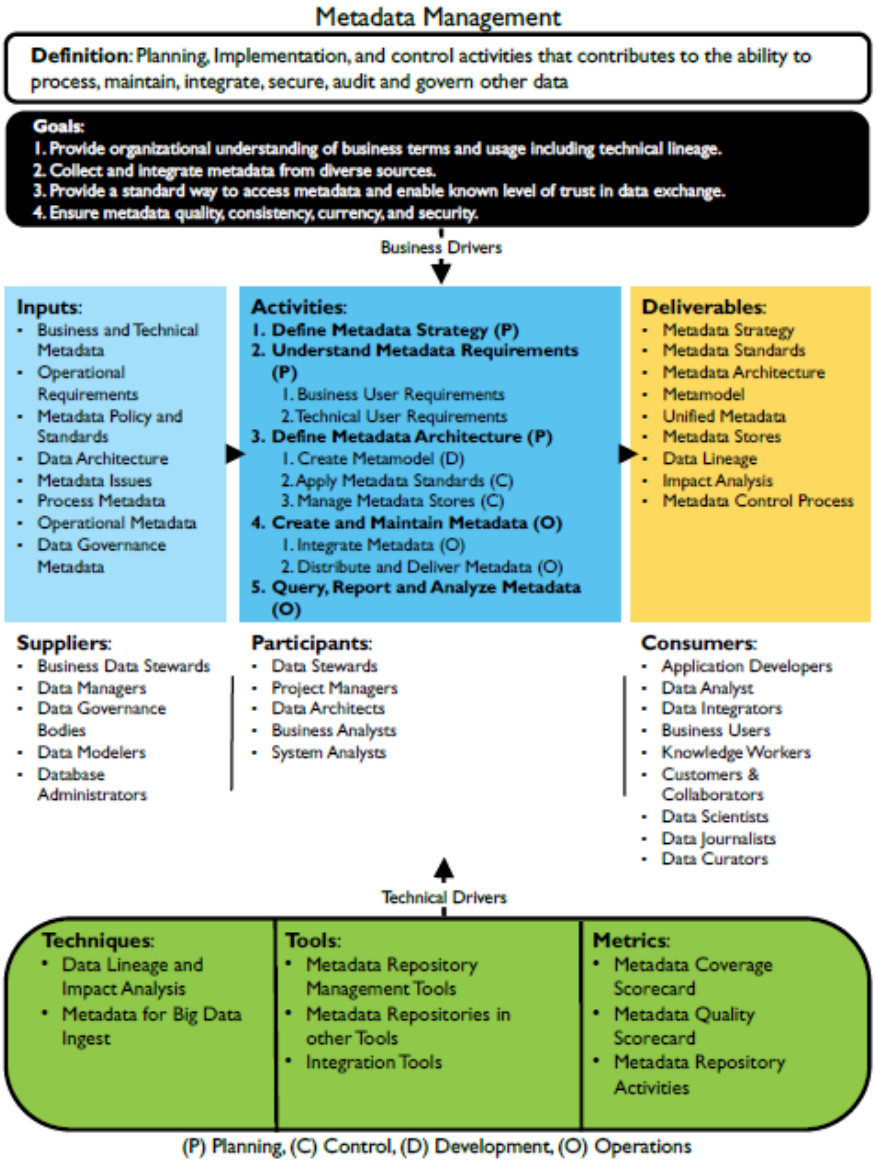


Figure 84 Context Diagram: Metadata

https://www.youtube.com/watch?v=HXAstVP3-y0

https://www.youtube.com/watch?v=Ioxr2mcOLjE

1.1 Business Drivers

Data cannot be managed without Metadata, which must also be managed. Metadata helps:

- Increase confidence in data by providing context, and measurement of data quality
- Increase value of strategic information (Master Data) by enabling multiple uses
- Operational efficiency by identifying redundant data and processes
- Prevent the use of out of date or incorrect data
- Reduce data-oriented research time
- Improve communication between business and IT
- Create accurate impact analysis
- Reduce system development lifecycle time
- Support regulatory compliance

1.2 Goals and Principles

Goals:
1. Provide organizational understanding of business terms and usage including technical lineage.
2. Collect and integrate metadata from diverse sources.
3. Provide a standard way to access metadata and enable known level of trust in data exchange.
4. Ensure metadata quality, consistency, currency, and security.

Implementation of a Metadata solution depends on the following principles:

- **Organisational commitment:** Senior manager support. Metadata management is an enterprise program
- **Strategy:** How Metadata will be created, maintained, integrated and accessed. Metadata strategy must align with business priorities
- **Enterprise perspective:** Ensure future extensibility but implement iteratively and incrementally
- **Socialisation:** Communicate the necessity and purpose of each type of Metadata
- **Access:** Ensure staff members know how to access and use Metadata
- **Quality:** Process owners accountable for quality of Metadata produced
- **Audit:** Set, enforce and audit standards for Metadata
- **Improvement:** Feedback mechanism for consumers

1.3 Essential Concepts

1.3.1 Metadata vs. Data

Organisations should define Metadata requirements and what they need it for. What is data and what is Metadata depends on the organisation.

Metadata is the "Who, What, Where, Why, When & How" of Data

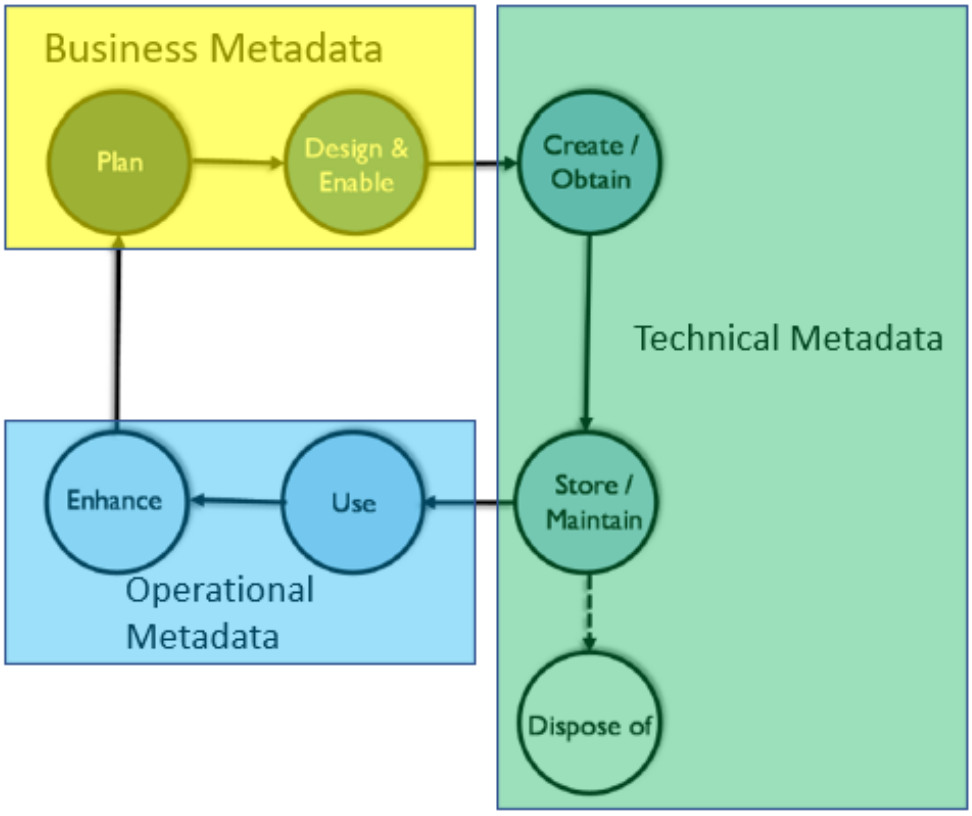
Who	What	Where	Why	When	How
Who created this data?	What is the business definition of this data element?	Where is this data stored?	Why are we storing this data?	When was this data created?	How is this data formatted? (character, numeric, etc.)
Who is the Steward of this data?	What are the business rules for this data?	Where did this data come from?	What is its usage & purpose?	When was this data last updated?	How many databases or data sources store this data?
Who is using this data?	What is the security level or privacy level of this data?	Where is this data used & shared?	What are the business drivers for using this data?	How long should it be stored?	
Who "owns" this data?	What is the abbreviation or acronym for this data element?	Where is the backup for this data?		When does it need to be purged/deleted?	
Who is regulating or auditing this data?	What are the technical naming standards for database implementation?	Are there regional privacy or security policies that regulate this data?			

Global Data Strategy, Ltd. 2017

7

1.3.2 Types of Metadata

(NB If an exam question has any other type, refer to the Appendix DMBOK V1 notes)



Lifecycle Phases and the Types of Metadata they Produce

Types:

- **Business Metadata:** Focus on the content and condition of data, and includes details related to Data Governance. Examples:
  - Definitions and descriptions of data sets, tables and columns
  - Business rules, transformation rules, calculations, and derivations
  - Data models
  - Data quality rules and measurement results
  - Schedules by which data is updated
  - Data provenance and data lineage
  - Data standards
  - Designations of the system of record for data elements
  - Valid value constraints
  - Stakeholders contact information (e.g. Data owners, data stewards)
  - Security/privacy level of data
  - Known issues with data
  - Data usage notes
- **Technical Metadata:** Provides information about technical details of data, systems that store it and the processes that move it. Examples:
  - Physical database table and column names
  - Column properties
  - Database object properties
  - Access permissions
  - Data CRUD rules
  - Physical data models, including table names, keys and the physical assets
  - Documented relationships between data models and physical assets
  - ETL job details
  - File format schema definitions
  - Source-to-target mapping documentation
  - Data lineage documentation, including upstream and downstream change impact information
  - Program and application names and descriptions
  - Content update cycle job schedules and dependencies
  - Recovery and backup rules
  - Data access rights, groups
- **Operational Metadata:** Describes details of the processing and accessing of data. Examples:
  - Logs of job execution of batch jobs
  - History of extracts and results
  - Schedule anomalies
  - Results of audit, balance, control measurements
  - Error logs
  - Reports and query access patterns, frequency, and execution time
  - Patches and Version maintenance plan and execution, current patching level
  - Backup, retention, date created, disaster recovery provisions
  - SLA requirements and provisions
  - Volumetric and usage patterns
  - Data archiving and retention rules, related archives
  - Purge criteria

- Data sharing rules and agreements
- Technical roles and responsibilities, contacts

1.3.3 ISO / IEC 11179 Metadata Registry Standard

ISO /IEC 11179 provides a framework for defining a Metadata registry, designed to enable Metadata driven data exchange based on exact definitions of data.

- Part 1: Framework for the Generation and Standardisation of Data Elements
- Part 3: Basic Attributes of Data Elements
- Part 4: Rules and Guidelines for the Formulation of Data Elements
- Part 5: Naming and Identification Principles for Data Elements
- Part 6: Registration of Data Elements

1.3.4 Metadata for Unstructured data

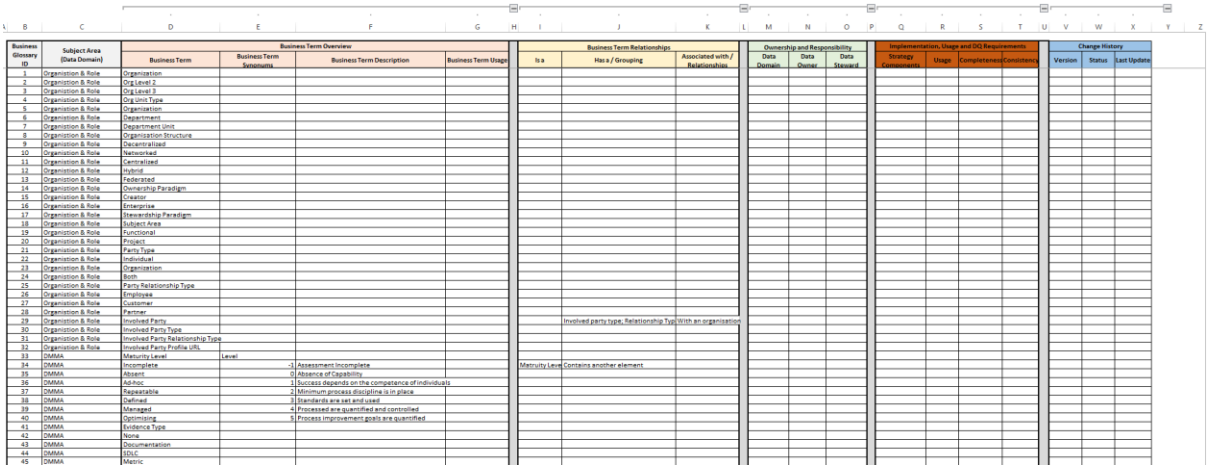
Types of Metadata for unstructured data (Stored with the document) – Documents and Big Data:

- **Descriptive:** Catalogue information and thesauri keywords
- **Structural:** Tags, field structures, format, terms on the Business Glossary
- **Administrative:** Sources, update schedules, access rights, navigation information
- **Bibliographic:** Library catalogue entries
- **Record keeping Metadata:** Retention policies
- **Preservation Metadata:** Storage, archival condition, rules for conservation, e-discovery

1.3.5 Sources of Metadata

- **Application Metadata Repositories:** The physical tables where Metadata is stored
- **Business Glossary:**
  - A document of the organisation's business concepts and terminology, definitions and the relationships between those terms. Accounts for hardware, software, database, processes and different user roles and responsibilities:
    - **Business users:** Data analysts, research analysts, management to understand terminology
    - **Data Stewards:** Manage the lifecycle of terms and definitions
    - **Technical users:** Use glossary terms to make design decisions
  - Business Glossary should capture business terms attributes such as:
    - Term name, definition
    - Ownership and stewardship
- **Business Intelligence (BI) Tools:** Overview information, classes, objects, derived and calculated items
- **Configuration Management Tools or Databases:** CMDB manage Metadata related to IT assets. Each asset is a Configuration Item (CI)
- **Data Dictionaries:** Defines the contents and structure of data sets. This Metadata is embedded in database/modelling tools, and must be extracted to use
- **Data Integration Tools:** Lineage Metadata, movement of data
- **Database Management and System Catalogues:** Describe the content, sizing information, software version, deployment status, network and infrastructure uptime, availability etc.
- **Data Mapping Management Tools:** Mapping tools and data integration tools can exchange data with Metadata repositories
- **Data Quality Tools:** Can share quality scores and patterns with Metadata repositories

- **Directories and Catalogues:** Contains information about systems, sources and locations of data in the organisation. It is useful for developers and super users
- **Event Messaging Tools:** Require a lot of Metadata to move messages between diverse systems
- **Modelling Tools and Repositories:** Produce Metadata relevant to the design of the application or system model
- **Service Registries:** Technical information about services and service end points e.g., APIs
- **Other Metadata Stores:** Specialised lists, repositories of repositories and business rules

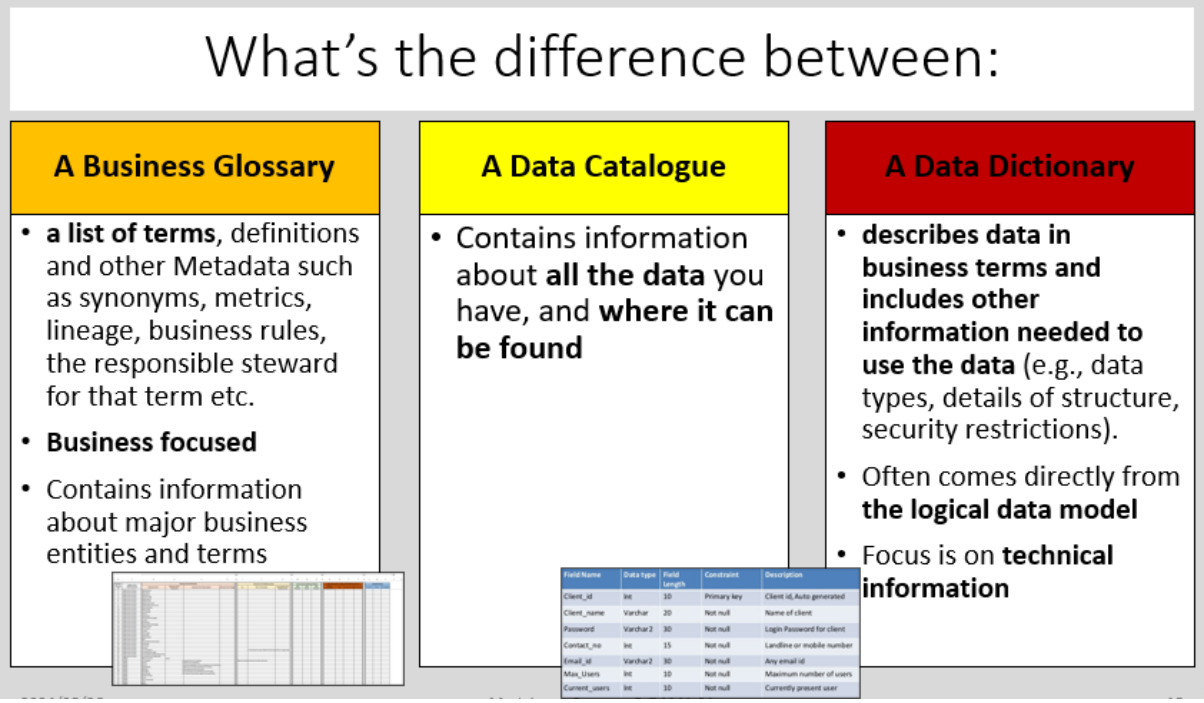


Business Glossary ID	Subject Area (Data Domain)	Business Term Overview				Business Term Relationships			Ownership and Responsibility			Implementation, Usage and DQ Requirements				Change History		
		Business Term	Business Term Synonyms	Business Term Description	Business Term Usage	Is a	Has a / Grouping	Associated with / Relationships	Data Domain	Data Owner	Data Steward	Strategy Comments	Usage	Completeness	Consistency	Version	Status	Last Update
1	Organistion & Role	Organization																
2	Organistion & Role	Org Level 2																
3	Organistion & Role	Org Level 3																
4	Organistion & Role	Org Unit Type																
5	Organistion & Role	Organization																
6	Organistion & Role	Department																
7	Organistion & Role	Department Unit																
8	Organistion & Role	Organization Structure																
9	Organistion & Role	Decentralised																
10	Organistion & Role	Networked																
11	Organistion & Role	Centralized																
12	Organistion & Role	Hybrid																
13	Organistion & Role	Federated																
14	Organistion & Role	Ownership Paradigm																
15	Organistion & Role	Creator																
16	Organistion & Role	Enterprise																
17	Organistion & Role	Stewardship Paradigm																
18	Organistion & Role	Subject Area																
19	Organistion & Role	Functional																
20	Organistion & Role	Project																
21	Organistion & Role	Party Type																
22	Organistion & Role	Individual																
23	Organistion & Role	Organisation																
24	Organistion & Role	Both																
25	Organistion & Role	Party Relationship Type																
26	Organistion & Role	Employee																
27	Organistion & Role	Customer																
28	Organistion & Role	Partner																
29	Organistion & Role	Involved Party					Involved party type; Relationship Type	With an organisation										
30	Organistion & Role	Involved Party Type																
31	Organistion & Role	Involved Party Relationship Type																
32	Organistion & Role	Involved Party Profile URL																
33	DMMA	Maturity Level	Level															
34	DMMA	Incomplete	-1	Assessment Incomplete		Matruity Level	Contains another element											
35	DMMA	Absent	0	Absence of Capability														
36	DMMA	Ad-hoc	1	Success depends on the competence of individuals														
37	DMMA	Repeatable	2	Minimum process discipline is in place														
38	DMMA	Defined	3	Standards are set and used														
39	DMMA	Managed	4	Processed are quantified and controlled														
40	DMMA	Optimizing	5	Process improvement goals are quantified														
41	DMMA	Evidence Type																
42	DMMA	None																
43	DMMA	Documentation																
44	DMMA	SDLC																
45	DMMA	Matric																

Excel Template of a Business Glossary from the Course How-to Govern Data by Howard Diesel

What to manage in a Business Glossary? (Not from DMBOK – Author Unknown)

- Term name, definition, acronym or abbreviation, and any synonyms
- Business unit and or application responsible for managing the data associated with the terminology
- Name of the person identifying the term, and date updated
- Categorization or taxonomy association for the term (business functional association)
- Conflicting definitions that need resolution, nature of the problem, action timeline
- Common misunderstandings in terms
- Algorithms supporting definitions
- Lineage
- Official or authoritative source for the data supporting the term


What's the difference between:
A Business Glossary
• a list of terms, definitions and other Metadata such as synonyms, metrics, lineage, business rules, the responsible steward for that term etc.
• Business focused
• Contains information about major business entities and terms
A Data Catalogue
• Contains information about all the data you have, and where it can be found
A Data Dictionary
• describes data in business terms and includes other information needed to use the data (e.g., data types, details of structure, security restrictions).
• Often comes directly from the logical data model
• Focus is on technical information

1.3.6 Types of Metadata Architecture

All Metadata management solutions include architectural layers which correspond to points in the Metadata lifecycle:

- Metadata creation and sourcing
- Metadata storage in one or more repositories
- Metadata integration: Bring together Metadata from various repositories
- Metadata delivery: Formats for delivery to another system
- Metadata usage: Searching
- Metadata control and management

NOTE: The information about the changes in the Revised Edition published on DAMA.ORG proposed the following changes:

- **Fig 85 remove arrows from Metadata to tools layer**
- **Fig 87 (Hybrid) add arrows from Metadata layer to tools layer**
- **Relabel fig 87**

NOTE: The arrow changes were NOT implemented. The small diagram is what appears in the DMBOK, the large diagram is the proposed change.

1.3.6.1 Centralised Metadata Architecture

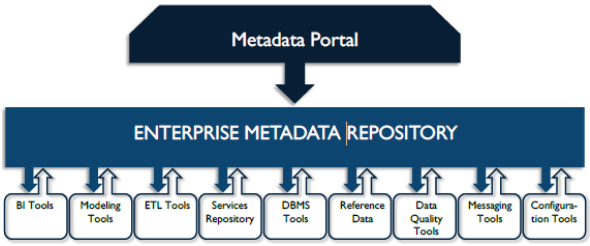
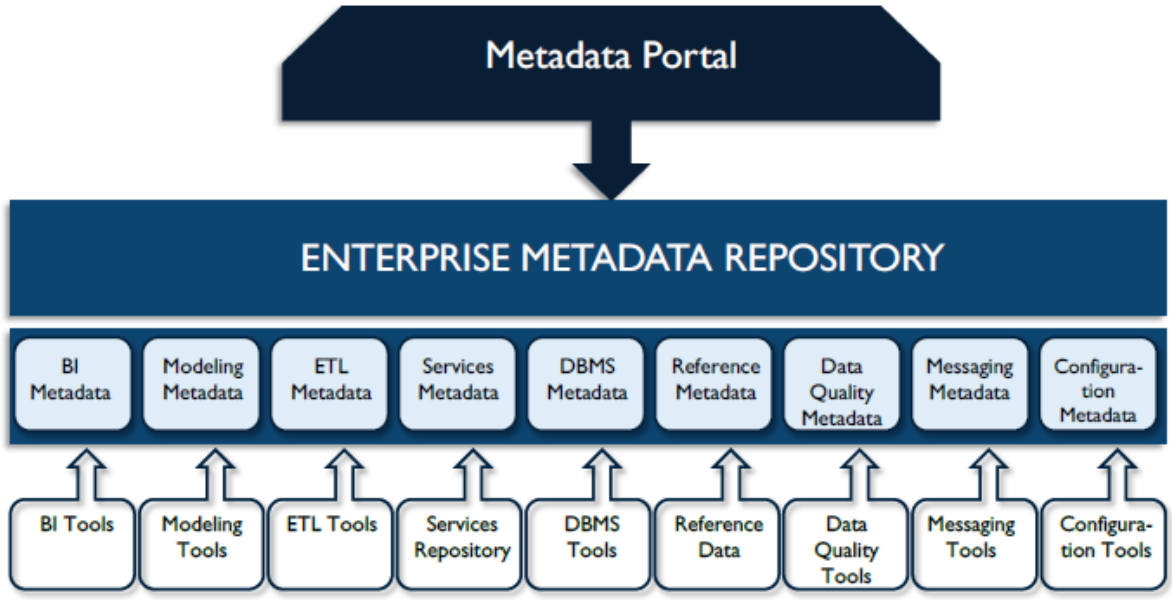


Figure 85 Centralised Metadata Architecture (Really the Bi-Directional)



A single Metadata repository containing copies of Metadata from the various sources.

Advantages of a centralised repository:

- High availability – independent of source systems
- Quick retrieval as repository and query resides together
- Resolved database structure not affected by proprietary nature of third party systems
- Quality is improved as extracted Metadata may be enhanced with Metadata from elsewhere

Limitations of a centralised repository:

- Complex processes ensure changes to source Metadata are quickly replicated into the repository
- Maintenance can be costly
- Extraction may require custom modules or middleware
- Increased demands on IT and vendor staff to maintain customised code

1.3.6.2 Distributed Metadata Architecture

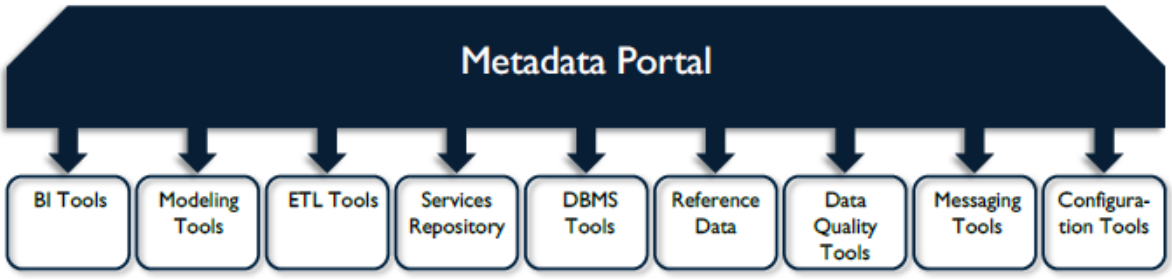


Figure 86 Distributed Metadata Architecture

No persistent repository, the portal passes user requests to the appropriate tool to execute. The Metadata retrieval engine retrieves data from source systems in real time. The Metadata management environment maintains source system catalogs and lookup information.

Advantages of the distributed Metadata architecture:

- Metadata is current and valid as it is retrieved from the source

- Queries are distributed – improved process and response time
- Implementation and maintenance effort minimised as Metadata requests from proprietary systems are queries, so no understanding of proprietary data structures is required
- Automated Metadata query processing is simpler
- Reduced batch processing

Limitations of distributed architecture:

- Not able to support user-added Metadata entries as there is no repository to put them
- Standardisation of presenting Metadata from various systems
- Query capabilities depend on the availability of the source systems
- Quality of Metadata depends on the source systems

1.3.6.3 Hybrid Metadata Architecture

Combines the Centralised and Distributed. The repository design only accounts for the user-added Metadata, critical standardised items and the additions from manual sources. Near real-time retrieval of Metadata from source and enhanced Metadata when needed.

1.3.6.4 Bi-Directional Metadata Architecture

Allows Metadata to change in any part of the architecture and then feedback is coordinated from the repository to the original source. The repository is forced to contain the latest version of the Metadata source, and must manage changes to the source as well.

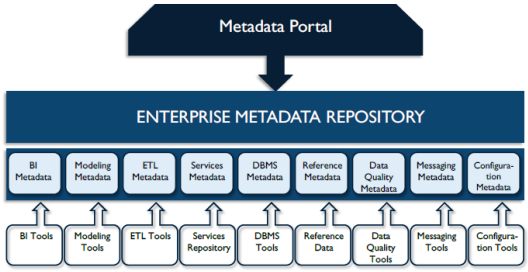
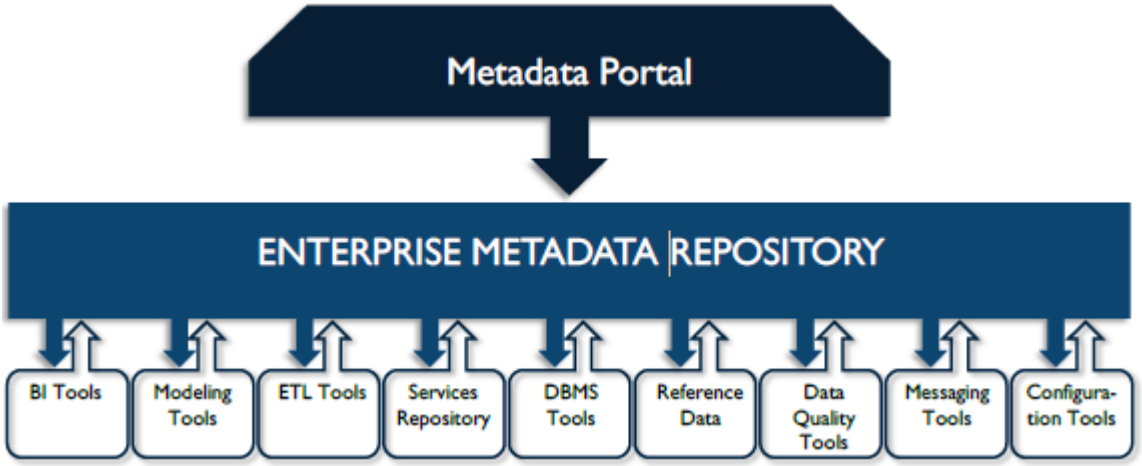


Figure 87 Bi-Directional Metadata Architecture (Really the Centralised)



2 Activities

2.1 Define Metadata Strategy

Define the future state enterprise Metadata architecture and the implementation phases:

- **Initiate Metadata strategy planning:** Key stakeholders involved in planning
  - Short- and long-term goals
  - Charter, scope, objectives and communications plan
- **Conduct key stakeholder interviews:** To provide foundational knowledge
- **Assess existing Metadata sources and information architecture:** Assess difficulty of project
  - Interview IT staff
  - Review system architecture and model documentation
- **Develop future Metadata architecture:** Develop long term architecture for the managed Metadata environment
- **Develop a phased implementation plan:** Prioritise findings from the interviews and data analyses to define a phased implementation plan

2.2 Understand Metadata Requirements

What Metadata is needed and at what level. Functionality-focussed requirements:

- **Volatility:** Update frequency
- **Synchronisation:** Timing of updates in relation to source changes
- **History:** Do historical versions need to be retained?
- **Access rights:** Who, how and interface functionality
- **Structure:** Metadata model for storage
- **Integration:** Degree of and rules for integration
- **Maintenance:** Processes and rules for updating
- **Management:** Roles and responsibilities
- **Quality:** Metadata quality requirement
- **Security:** Some Metadata cannot be exposed as it will reveal sensitive data

2.3 Define Metadata Architecture

Metadata Management System must be able extract Metadata from many sources by scanning the sources and updating the repository, while supporting manual updates, searches and lookups by various user groups. Single access point for the Metadata repository which is transparent to users.

2.3.1 Create MetaModel

Data model for the Metadata repository:

- High-level conceptual model – relationships between systems
- Lower level metamodel that describes the elements and processes
- The metamodel is a planning tool, and Metadata itself

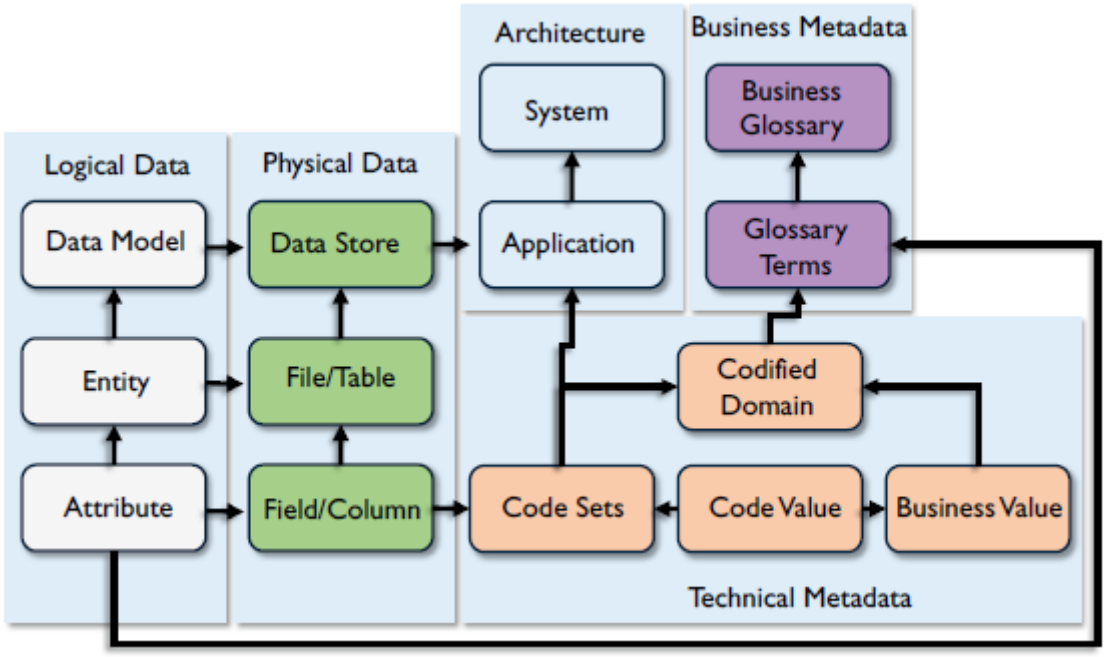


Figure 88 Example Metadata Repository Metamodel

2.3.2 Apply Metadata Standards

Governance monitors for compliance:

- Internal standards such as naming conventions (ISO 11179 for naming conventions)
- External standards such as data exchange formats

2.3.3 Manage Metadata Stores

Control activities should have governance oversight and include:

- Job scheduling and monitoring
- Load statistical analysis
- Backup, recovery, archive, purging
- Configuration modifications
- Performance tuning
- Query statistics analysis
- Query and report generation
- Security management

Quality control activities:

- QA, quality control
- Matching update sets to timeframes
- Missing Metadata reports

Metadata Management activities include:

- Loading, scanning, importing and tagging assets
- Source mapping and movement
- Versioning
- User interface management
- Linking data sets Metadata maintenance – for NoSQL provisioning

- Linking data sets to internal data acquisition – custom links and job Metadata
- Licensing for external data sources and feeds
- Data enhancement Metadata e.g., Link to GIS

Training:

- Education and training of users and data stewards
- Management metrics generation and analysis
- Training on the control activities and query and reporting

2.4 Create and Maintain Metadata

Metadata should be planned and created as a product. Profile and inspect for quality. Schedule maintenance.

General principles of Metadata management:

- **Accountability:** Metadata is often produced through existing processes – hold those process owners accountable for quality of Metadata
- **Standards:** Set, enforce and audit Metadata standards
- **Improvement:** Create a consumer feedback mechanism

2.4.1 Integrate Metadata

As Metadata is gathered from many sources, and integrated into the Metadata repository, challenges arise which require governance.

Two approaches for repository scanning:

- **Proprietary interface:** Collection and loading of Metadata occurs in single step. No format specific file output
- **Semi-proprietary interface:** Two step process where scanner collects Metadata from a source and outputs it to a format specific data file

Files used during the scanning process:

- **Control file:** Contains the source structure of the data model
- **Reuse file:** Contains the rules for managing reuse of process tools
- **Log file:** Produced during each phase of the process
- **Temporary and backup files:** Used for traceability

2.4.2 Distribute and Deliver Metadata

Metadata is delivered to consumers/applications/tools requiring Metadata feeds:

- Metadata intranet websites
- Reports, glossaries and other documents
- Data warehouses, data marts and BI tools
- Modelling and software development tools
- Messaging and transactions
- Web services and APIs
- External organisation interface solutions

Metadata is exchanged with external organisations using files (flat, XML or JSON) or web services.

2.5 Query, Report and Analyse Metadata

A Metadata repository has a front-end application for search and retrieval as Metadata guides the use of data assets.

3 Tools

The Metadata repository is the primary tool used to manage Metadata. It has an integration layer and an interface for manual updates. Metadata repository management tools are also a source of Metadata.

4 Techniques

4.1 Data Lineage and Impact Analysis

Metadata about the physical assets provides information about how the data is transformed as it moves between systems. Limited to the scope of the Metadata management system.

- **'As Implemented Lineage'**:
  - Current version of lineage based on programming code
  - Imported from various tools
- **'As Designed Lineage'**:
  - Lineage described in mapping specification documents
  - Not extractable by Metadata management system
- **Stitching**: The process whereby the Metadata management system augments the 'As Implemented' data lineage with the 'As Designed'

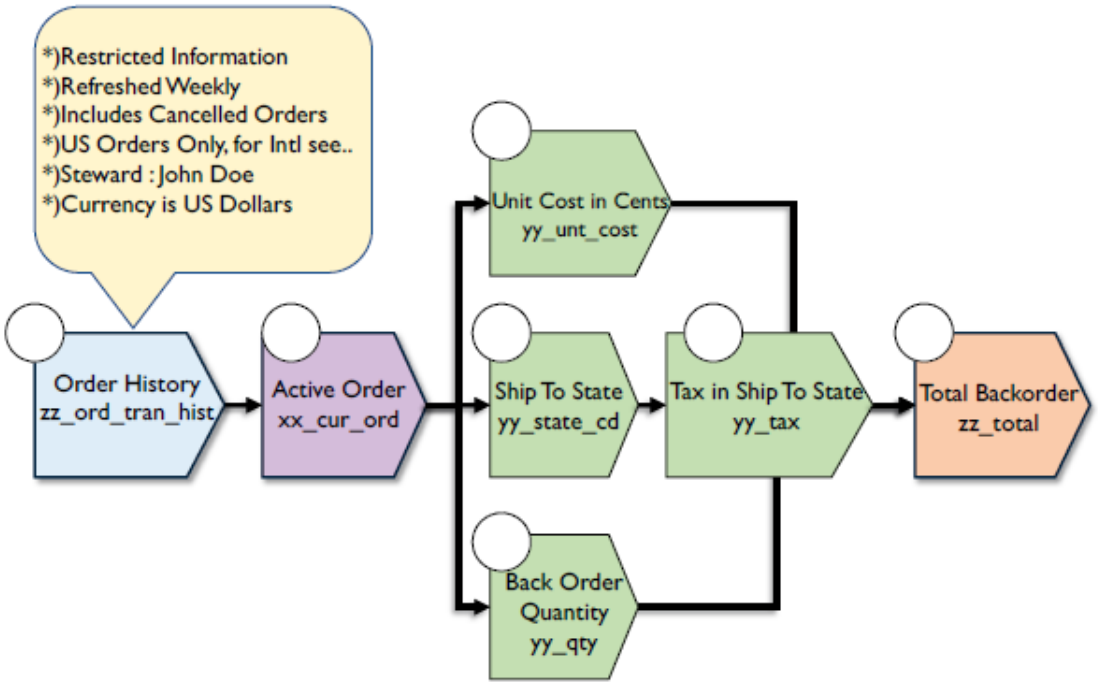


Figure 89 Sample Data Element Lineage Flow Diagram

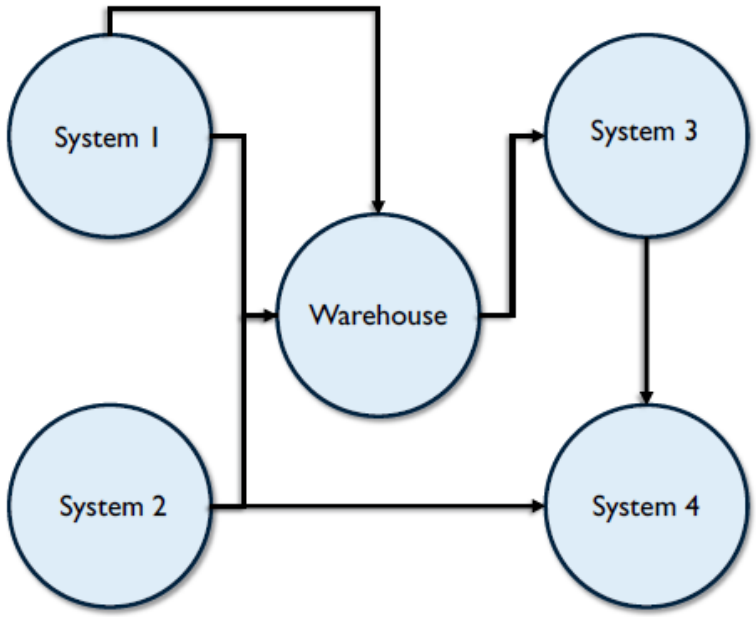


Figure 90 Sample System Lineage Flow Diagram

Successful lineage discovery needs to account for both business and technical focus:

- **Business focus:** Data elements prioritised by business
  - Start at target and trace back to source
  - Gives business understanding what happens to a data element as it moves
  - DQ measurements with lineage can pinpoint where system design impacts quality
- **Technical focus:**
  - Start at source systems
  - Identify immediate consumers, then next sets of consumers until all systems are identified

4.2 Metadata for Big Data Ingest

Metadata tags should be applied to data on ingestion to the data lake. Ingestion engines can profile data as it is ingested to identify data domains, relationships and DQ issues. PPI and sensitive data can be tagged, as well as codes and confidence tags for data science.

5 Implementation Guidelines

Implement the Metadata environment in incremental steps to minimise risks and facilitate acceptance. Use a relational database platform. Contents should be generic in design and should be integrated so that consumers can see across different data sources. Should house current, planned and historical versions of Metadata.

5.1 Readiness assessment / Risk Assessment

People should be aware of the risks of not managing Metadata:

- Errors in judgement due to lack of knowledge of the context of data
- Exposure of sensitive data
- Risk that SMEs will leave and take their knowledge of the data with them

A formal assessment of the current maturity of Metadata activities includes:

- Critical data elements
- Available Metadata glossaries
- Lineage
- Data profiling and data quality processes
- MDM maturity

5.2 Organisational and Cultural Change

Metadata efforts often meet with resistance. Needs senior management support and engagement. Business and technical staff work closely in a cross-functional manner.

6 Metadata Governance

Determine the specific requirements for the management of the Metadata lifecycle, and establish governance processes. Formal roles and responsibilities need to be assigned to dedicated resources.

6.1 Process Controls

Governance team responsible for:

- Defining standards
- Managing status changes for Metadata
- Promotion of Metadata
- Training
- Management of business terms

Metadata strategy should be integrated into the SDLC to ensure Metadata is collected and remains current.

6.2 Documentation of Metadata Solutions

A master catalogue of Metadata of the sources and targets currently on scope. It is a 'what-is-where' guide for the user community and includes:

- Metadata implementation status
- Source and target Metadata store
- Schedule information for updates
- Retention and versions kept
- Contents
- Quality statements or warnings
- System of record or other data source statuses
- Tools, architecture and people involved
- Sensitive information and removal or masking for the source

6.3 Metadata Standards and Guidelines

Metadata standards are required in the exchange of data with operational trading partners.

- Use industry-based Metadata standards early
- Tool vendors provide XML, JSON or REST support to exchange their data
- Tools offer import/export capabilities using XML
- Templates and examples
- ISO standards

6.4 Metrics

- **Metadata repository completeness:** Ideal coverage compared to actual coverage
- **Metadata Management Maturity:** Based on the Capability Maturity Model (CMM-DMM) assessment approach
- **Steward representation:** Coverage across enterprise for stewardship
- **Metadata usage:** User uptake measured in logins
- **Business Glossary activity:** Usage, update, resolution of definitions, coverage
- **Metadata documentation quality:** Assess automatically and manually
- **Metadata repository availability:** Uptime, processing time (batch and query)

Chapter 13
Data Quality Management

1 Introduction

Data Quality is the degree to which dimensions of Data Quality meet the requirements; therefore, high quality of data is seen in context as "fit for purpose".

The Data Quality Function includes programs, projects and ongoing operational work. Long term success requires the organisation to adopt a quality mindset, driven by people who use data.

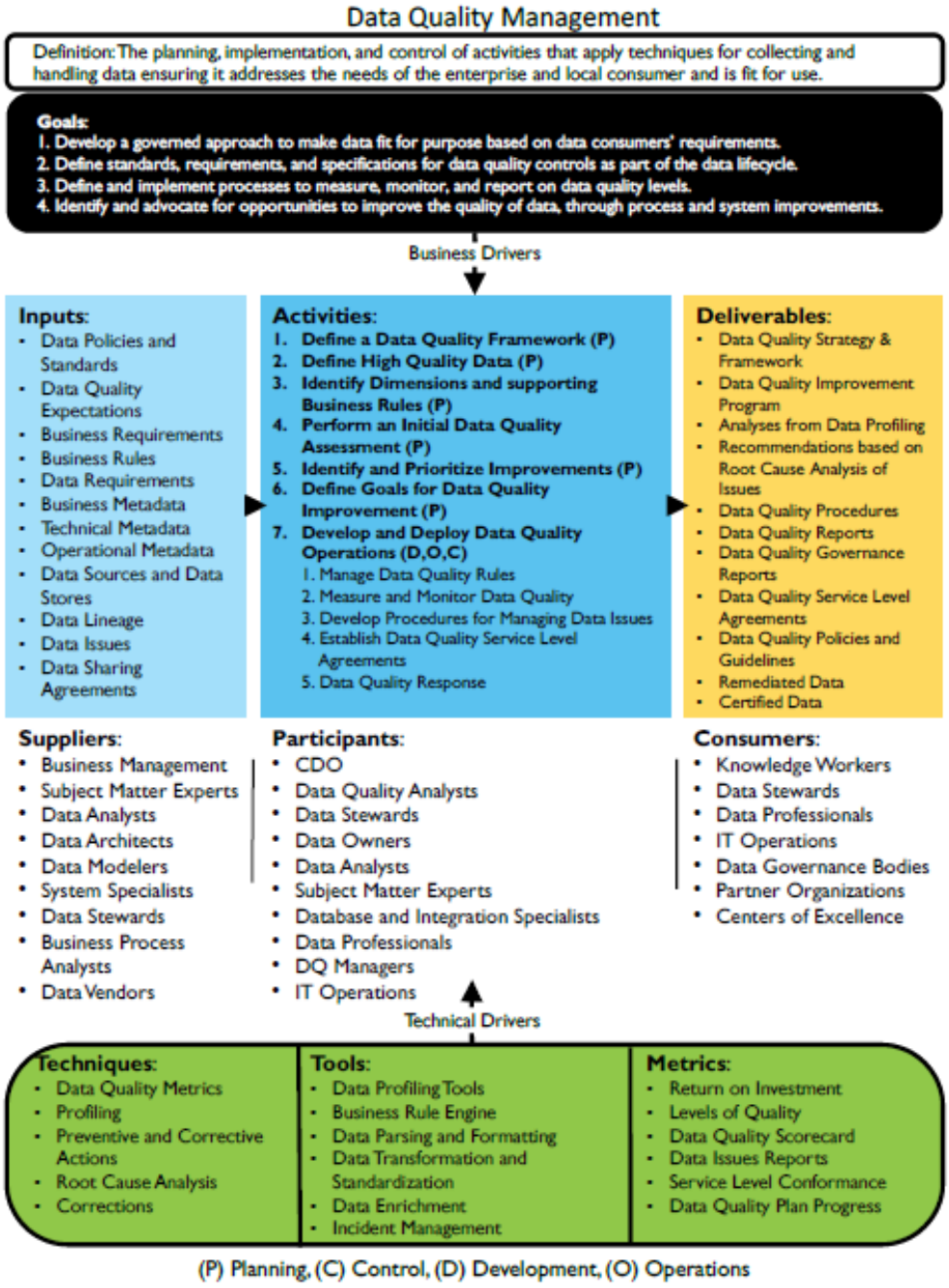

Data Quality Management
Definition: The planning, implementation, and control of activities that apply techniques for collecting and handling data ensuring it addresses the needs of the enterprise and local consumer and is fit for use.
Goals:
1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.
Business Drivers
Inputs:
Data Policies and Standards
Data Quality Expectations
Business Requirements
Business Rules
Data Requirements
Business Metadata
Technical Metadata
Operational Metadata
Data Sources and Data Stores
Data Lineage
Data Issues
Data Sharing Agreements
Activities:
1. Define a Data Quality Framework (P)
2. Define High Quality Data (P)
3. Identify Dimensions and supporting Business Rules (P)
4. Perform an Initial Data Quality Assessment (P)
5. Identify and Prioritize Improvements (P)
6. Define Goals for Data Quality Improvement (P)
7. Develop and Deploy Data Quality Operations (D,O,C)
1. Manage Data Quality Rules
2. Measure and Monitor Data Quality
3. Develop Procedures for Managing Data Issues
4. Establish Data Quality Service Level Agreements
5. Data Quality Response
Deliverables:
Data Quality Strategy & Framework
Data Quality Improvement Program
Analyses from Data Profiling
Recommendations based on Root Cause Analysis of Issues
Data Quality Procedures
Data Quality Reports
Data Quality Governance Reports
Data Quality Service Level Agreements
Data Quality Policies and Guidelines
Remediated Data
Certified Data
Suppliers:
Business Management
Subject Matter Experts
Data Analysts
Data Architects
Data Modelers
System Specialists
Data Stewards
Business Process Analysts
Data Vendors
Participants:
CDO
Data Quality Analysts
Data Stewards
Data Owners
Data Analysts
Subject Matter Experts
Database and Integration Specialists
Data Professionals
DQ Managers
IT Operations
Consumers:
Knowledge Workers
Data Stewards
Data Professionals
IT Operations
Data Governance Bodies
Partner Organizations
Centers of Excellence
Technical Drivers
Techniques:
Data Quality Metrics
Profiling
Preventive and Corrective Actions
Root Cause Analysis
Corrections
Tools:
Data Profiling Tools
Business Rule Engine
Data Parsing and Formatting
Data Transformation and Standardization
Data Enrichment
Incident Management
Metrics:
Return on Investment
Levels of Quality
Data Quality Scorecard
Data Issues Reports
Service Level Conformance
Data Quality Plan Progress
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 91 Context Diagram: Data Quality

The value of data is that data is reliable and trustworthy i.e., of high quality.

Factors that contribute to poor quality data:

- Lack of understanding of the effects of poor quality data on the decision-making process
- Bad Planning
- Siloed system design
- Inconsistent development processes
- Incomplete documentation
- A lack of standards
- A lack of governance
- Failure to define what makes data fit for purpose

High quality data should be the goal of all data management disciplines and they all contribute to the quality of data. Data Quality should be managed by a Data Quality function team responsible for engaging both business and technical data management professionals to apply quality management techniques to data to ensure it is fit for a variety of purposes.

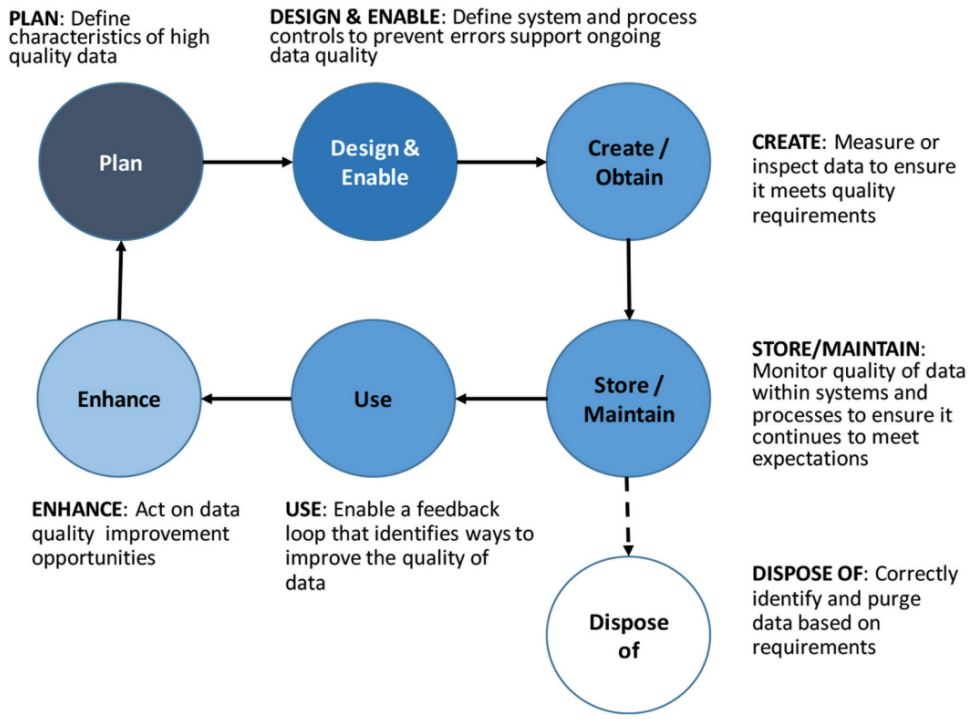

PLAN: Define characteristics of high quality data
DESIGN & ENABLE: Define system and process controls to prevent errors support ongoing data quality
Plan
Design & Enable
Create / Obtain
CREATE: Measure or inspect data to ensure it meets quality requirements
Enhance
Use
Store / Maintain
STORE/MAINTAIN: Monitor quality of data within systems and processes to ensure it continues to meet expectations
ENHANCE: Act on data quality improvement opportunities
USE: Enable a feedback loop that identifies ways to improve the quality of data
Dispose of
DISPOSE OF: Correctly identify and purge data based on requirements

Figure 27: Data Quality Management and the Data Lifecycle (Adapted from DMBOK2, p. 29)

From Navigating the Labyrinth by Laura Sebastian-Coleman

1.1 Business Drivers

Business drivers for establishing a Data Quality Program:

- Enhancing the stakeholder experience and the organisation's reputation
- Increasing the effectiveness of the organisation
- Reducing risks and costs associated with poor quality data
- Improving organisational efficiency and productivity

DQ management achieves these using a systematic approach by:

- Understanding the context driving the need for data (fitness)
- Matching the DQ needs to this context (purpose)
- Ensuring repeatability through stable process

Tangible benefits:

- Customers trusting the data and have confidence dealing with organisation
- Employees answer questions quickly and consistently
- Revenue gains through
  - Identifying business opportunities
  - Effective invoicing
- Decreased customer service calls
- Staff spending less time on figuring out if the data is right, and more time on insights
- Regulatory compliance and improved credit rating

1.2 Goals and Principles

General goals:

Goals:
1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.

A Data Quality Program should be guided by the following principles:

- **Criticality:** Focus on data most critical to the enterprise and its customers
- **Standards-driven**: Define stakeholders requirements as measurable standards
- **Objective measurement and transparency:** Measurement and methodology must be communicated to stakeholders
- **Prevention:** Focus should be on prevention of data errors
- **Root cause remediation:** Don't just correct errors, address the root cause, which may be poor process design
- **Embedded in business processes:** Business process owners must enforce data quality standards
- **Systematically enforced:** System owners must systematically enforce data quality requirements
- **Connected to service levels:** Data quality reporting and issues management must be incorporated into SLAs

1.3 Essential Concepts

1.3.1 Critical Data

Critical Data Elements are identified by linking them to the Data Quality Business Drivers. How to identify critical data. It is often used in:

- Regulatory, financial, or management reporting
- Business operational needs
- Measuring product quality and customer satisfaction
- Business strategy, especially efforts at competitive differentiation

- Master and Reference Data are critical by definition

1.3.2 Data Quality Dimensions

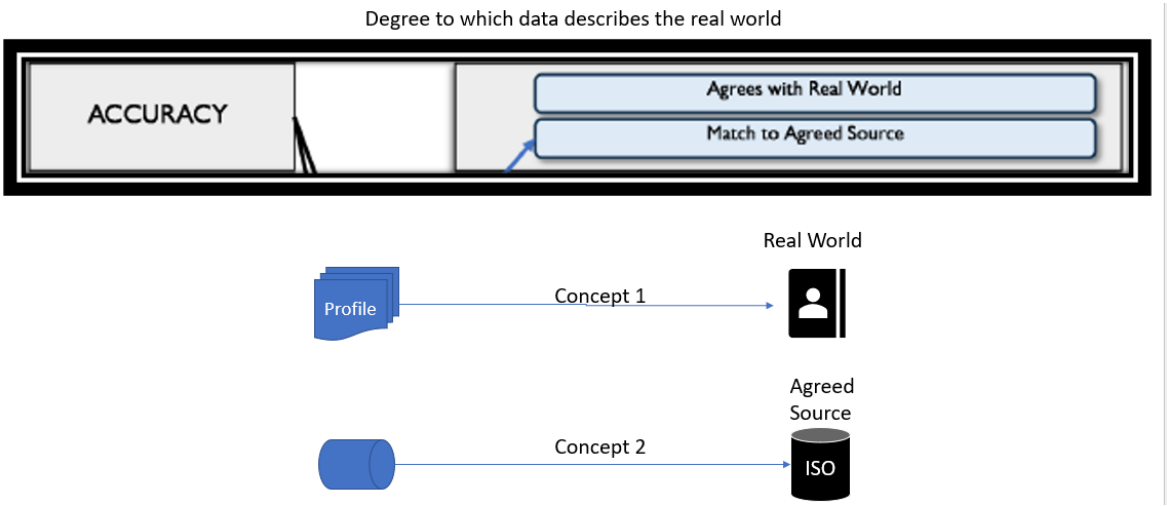
DQ Dimensions are measurable characteristics of data which form the basis for measurable rules.

Dimension of Quality	Description
Validity	Validity refers to whether data values are consistent with a defined domain of values. A domain of values may consist of: • A data type (e.g., text with or without special characters, a number, or a date) • A data range (e.g., a numeric range, date range, or a set of valid text values in a reference table) • A format (e.g., a phone number must contain country, area codes, and the correct number of digits or a currency must have only two decimal places) • The precision of expected values. For example, a date-stamp must be recorded to a millisecond or a number recorded to no more than two decimal places. • The time frame (certain values may also only be valid for a specific length of time, for example, data that is generated from RFID (radio frequency ID) or some scientific data sets).
Completeness	Completeness refers to whether all the required data is present. For example, are all the mandatory components of an address populated, including the street number, street name, city, and country? For example, if the customer account is marked as a debt risk, there must be a debt manager appointed. If the customer account is not marked as a debt risk, there must not be a debt manager appointed. Completeness can be measured at the column, record, or data set level. • Are columns/data elements populated to the level expected? (Some columns are mandatory. Optional columns are populated only under specific conditions.) • Are records populated correctly? (Records with different statuses may have different expectations for completeness.) • Does the data set contain all the records expected?
Consistency	Consistency is ensuring that data values are coded using the same approach, assessment and valuation criteria. Consistency is between two different data items, potentially within a data set and between data sets, and across time. Consistency may be defined between one set of data element values and another data element set within the same record (record-level consistency), between one set of data element values and another data element set in different records (cross-record consistency), or between one set of data element values and the same data element set within the same record at different points in time (temporal consistency). Consistency can also be used to refer to consistency of format. Take care not to confuse consistency with accuracy or correctness. Consistency can be measured using the following ways: • Cross record consistency is between data values in the same column – are all the customer addresses recording the head-office details or do some record service delivery centers? • Between associated data sets, where data sets are expected to have values that are linked. For example, when moving data from data capture systems to data warehousing systems, a consistency quality measure will maintain confidence in the values in the data warehouse. • Across time can be applied to any of the previous cases. For example, has the grading of students been the same over time? If a subject is completed previously, would the same learning performance yield the same grade today?
Integrity	Integrity refers to the lack of incoherent values and broken relationships in data. Integrity can be measured using the following ways: • Coherence – where one data value implies a limited range in another data value and they are matching. For example, a country has a defined set of states or provinces. If a customer address is in a certain country, they must use a state or province name for the associated set. For example, Alberta is a province of Canada, Pisa is a province of Italy, and Tasmania is a state of Australia. • Parent-child (referential integrity) – where every child must have a parent data value. For example, all the customer address country names must exist on the valid list of countries where the organization is authorized to sell products to. Data sets without integrity are seen as corrupted or have data loss. Data sets without referential integrity have 'orphans' – invalid reference keys, or 'duplicates' – identical rows which may negatively affect aggregation functions.

Dimension of Quality	Description
Timeliness	Data Timeliness refers to the time that data needs to become accessible to a user after its capture or update. Timeliness is the expectation and actual delay before the data becomes available. For example, An electricity utility needs to maintain its power grid and must have demand data measured and available to a system operator within a few seconds. Otherwise, electricity generation may get out of alignment with demand, causing an energy excess or shortage. As another example, a statistical agency of government produces the Gross Domestic Product (GDP) report two months after the end of the quarter. In order to meet this timeliness requirement, all the data collections that make up the GDP will have their own specific timeliness requirements.
Currency	Currency is the date that the data was last updated relative to now and the likelihood that it is still correct. Different data sets will have different currency expectations from relatively static data to highly volatile data. Static data remains current for a long period. Volatile data remains current for a relatively short period. For example, country codes are relatively static, remaining current for a long period. A business' bank account is relatively volatile as the balance is constantly changing. Volatile data, like stock prices on financial web pages, will often be shown with an "as-of-time", so that data consumers understand the risk that the data has changed since it was recorded. During the day, while the markets are open, such data will be updated frequently – it is volatile. Once markets close, the data will remain unchanged but will still be current since the market itself is inactive –it is static. Key terms used in currency include: • Update time – the time stamp of the last update • Volatility measures the rate of change of the date. It may measure across all the data values or measure within a segment. For example, the rate that domestic customers change their address may be different from the rate that corporate customers change their address. • Latency measures the time between when the data was created and when it was made available for use. For example, overnight population of a data warehouse may end up with different data latencies. Data can have a latency of 1 day for data entered into the system early on the prior day, but only a few minutes for data generated just before the load. (See **Chapter 8.**).
Reasonableness	Reasonableness asks whether a data pattern meets expectations. For example, whether a distribution of sales in a geographic area makes sense based on what is known about the customers in that area. For example, based on previous customer logins at 5pm, are today's customer logins to our systems out of the ordinary? Key terms used in reasonableness include: • Fixed values – this is where the measure is fixed, not dependent on previous behavior. The maximum currency value of a transaction may be limited in size by the technical limits of a bank's systems and other values bigger than this value is "unreasonable". • Benchmark value – it is possible to measure the average, deviation or any other statistical function and ensure the latest set of values conforms to expected boundaries. For example, there have been between 5,0000 and 10,000 customer address changes per day, but yesterday there were 2,000,000. On a government data collection, the average size company has 20 employees, but the latest returns had an unusual number with an average of 1,500 employees reported.
Uniqueness / Deduplication	Uniqueness states that no real-world entity exists more than once within the data set. Asserting uniqueness of the entities within a data set implies that each row relates to each unique entity in the real world, and only that specific entity. Uniqueness can be identified using the following techniques: • Key Structure – where there are duplicate keys within the data set (See Chapter 5). For example, in a data set containing customers, many of which share the same customer number. • Related data – where other data in the data may point to a duplicate. For example, two customers with different customer numbers, but having the same name, date of birth, and office address. • Inquiries or complaints that users cannot access certain data. For example, a customer with multiple services can only access some of their services as the others are recorded under a different customer number (which is a duplicate).

Dimension of Quality	Description
Accuracy	Accuracy refers to the degree that data correctly represents 'real-life' entities. For example, is the person's name in our database actually the person's name in real life? Does the customer actually use that email address? Accuracy is difficult to measure unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy use the other dimensions to imply accuracy. Common techniques to improve accuracy include: • Checking consistency with a data source that has been verified as accurate; a government company register (system of record) or a commercial company Information Broker (system of reference). • Ongoing calibration of devices with reality. For example, over time, weather gauges may drift from providing accurate measurements. An annual program of field inspections visits 10% of the devices and recalibrates them to a standard to ensure they are delivering accurate measurements. • Sampling data values with reality. For example, small number of emails are sent to customers (mainly) to drive sales, but also have the benefit of checking the accuracy of the email address (does it exist, is it used)?

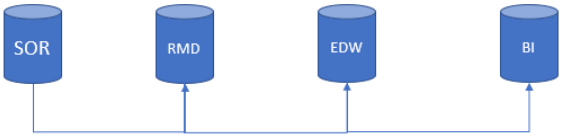
Table 29 Common Dimensions of Data Quality



Explanation of DQ Dimension Accuracy from Figure 92 (Howard Diesel)

Data values in one data set are consistent with values in another data set

Concept 1: Distributed Data



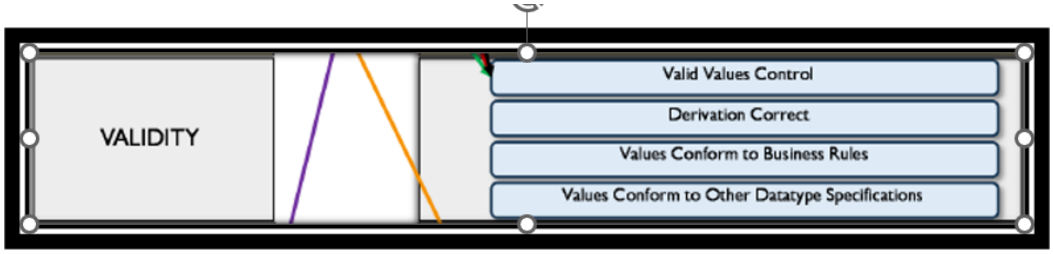
Concept 2: Format

Data

Credit Card	Consistency	Score	Expect
####-####-####-####		50%	70%
###-###	⊘		
####-####-####-####			
####-####-####	⊘		

Fail

Credit Card Number Business Rule
####-####-####-####



Concept 1: Valid Values Control

All State / Province names must be a value from the reference table stored in the Metadata repository

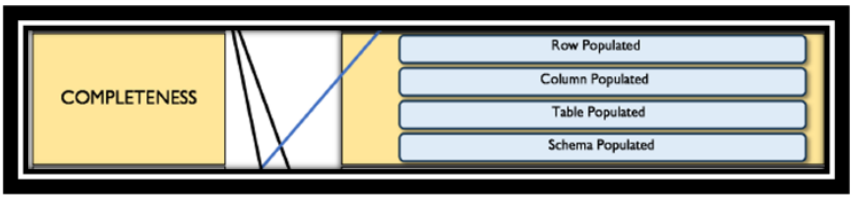
Concept 3: Values Conform to Business Rules

Credit Card Number Business Rule ####-####-####-####

Data

Credit Card	Validity	Score	Expect
####-####-####-####		50%	70%
###-###	⊘		
####-####-####-####			
####-####-####	⊘		

Fail



Certain attributes always have assigned values

Credit Card	Completeness	Score	Expect
	⊘	75%	90%
###-###			
####-####-####-####			
####-####-####			

Fail

Credit Card	Completeness	Score	Expect
####-####-####-####		100%	90%
###-###			
####-####-####-####			
####-####-####			

Pass

Explanation of DQ Dimension Completeness from Figure 92 (Howard Diesel)

Degree to which information is current / up to date

Age	Shoe Size	Last Update	Currency	Score	Expect
5	10	15 Jan 2015	⊘	50%	75%
10	2	15 Jan 2020	⊘		
15	8	15 Oct 2021			
30	10	15 Apr 2016			

Today: 8 Nov 2021

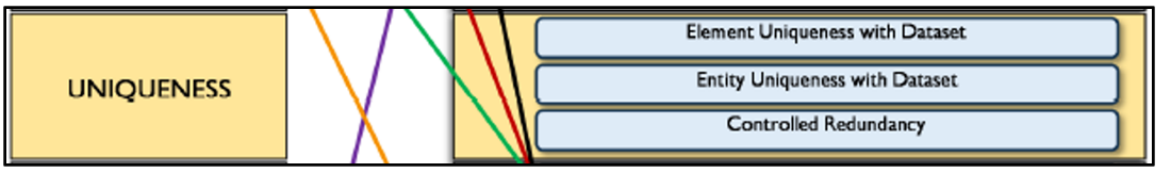
Fail

Currency Business Rule
Shoe Size
Age Band: 5 – 10
Currency: 6 Months
Age Band: 11 - 20
Currency: 1 year
Age Band: 21 – 50
Currency 10 year

Age	Shoe Size	Last Update	Currency	Score	Expect
5	10	15 May 2021		75%	75%
10	2	15 Jan 2020	⊘		
15	8	15 Oct 2021			
30	10	15 Apr 2016			

Pass

Explanation of DQ Dimension Currency from Figure 92 (Howard Diesel)



No entity instance recorded more than once

National ID	Uniqueness	Score	Expect
1234567890		75%	100%
0987654321			
1234567891			
1234567890	⊘		

Fail

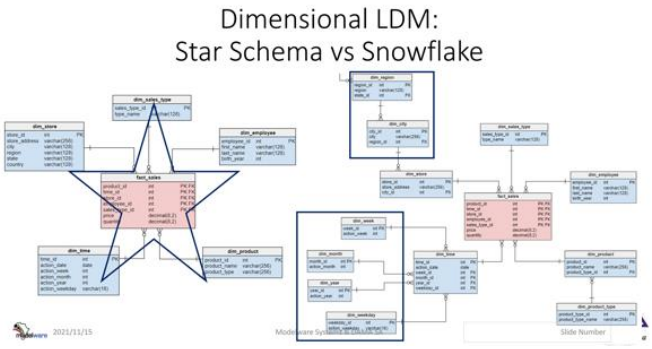
National ID	Uniqueness	Score	Expect
1234567890		100%	100%
0987654321			
1234567891			
1234567892			

Pass

Explanation of DQ Dimension Uniqueness from Figure 92 (Howard Diesel)

A **controlled redundancy** is a necessary technique to use redundant fields in a database. This speed ups the database access and also improves the performance of queries. Usually, the DBMS ensures the allocation of the data in the records. It should have the capability to control this redundancy in order to prohibit inconsistencies among the files.

Dimensional Technique:
Flatten Dimensions to Remove Snowflake in Star Schema



Explanation of DQ Dimension Uniqueness Concept Controlled Redundancy from Figure 92 (Howard Diesel)

Degree to which data represents reality at any point in time

Timeliness	Time Expectation for Availability	The measure of time between when data is expected versus made available.
	Manual Float	Manual float is a measure of the time from when an observation is made to the point it is recorded in electronic format.
	Electronic Float	Electronic float is a measure of the time from when data is captured in an electronic format until it is accessed by a person.

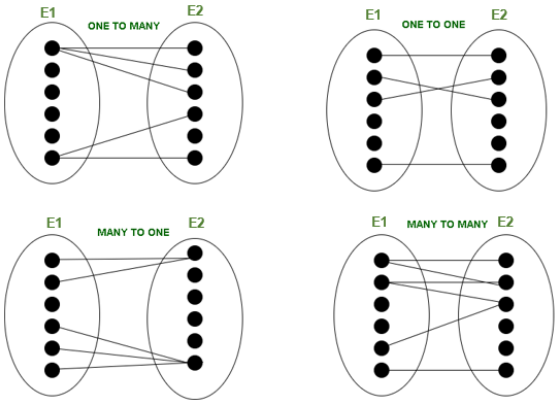
Explanation of DQ Dimension Timeliness from Figure 92 (Howard Diesel)

Concepts

1. Unique ID of Entity: Primary Key
2. Cardinality: Relationships
3. Referential Integrity: Foreign Key

Explanation of DQ Dimension Data Integrity from Figure 92 (Howard Diesel)

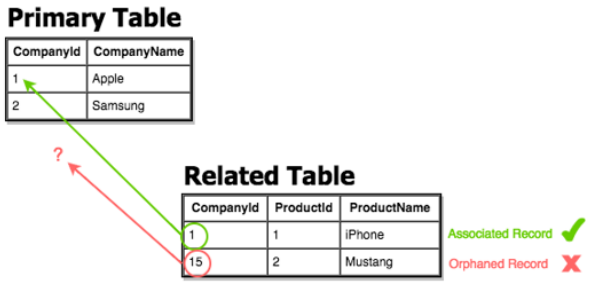
Integrity	Referential Integrity	Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
	Uniqueness	Uniqueness measures whether each fact is uniquely represented.
	Cardinality	Cardinality describes the relationship between one data set and another, such as one-to-one, one-to-many, or many-to-many.



Concept 3: Referential Integrity (Foreign Keys)

Referential integrity will prevent users from:

- Adding rows to a related table if there is no associated row in the primary table.
- Changing values in a primary table that result in orphaned records in a related table.
- Deleting rows from a primary table if there are matching related rows



Explanation of DQ Dimension Data Integrity from Figure 92 (Howard Diesel)

1.3.3 Data Quality Business Rules

Business Rules describe how data should exist to be useful to the organisation. Data Quality rules are aligned with DQ Dimensions to describe DQ requirements. Example, Business rules for State / Province names:

- **Validity:** Must be a value from a reference table in the Metadata repository
- **Completeness:** All addresses must have a state / province name if the country is divided
- **Integrity:** State / province names must be linked to the country
- **Currency:** Reference values must be periodically updated and address table brought up to date

Notes:

- Not every dimension is applicable every time
- Often business rules are compressed to cover multiple dimensions and data elements
- Implementation:
  - Data entry pick lists
  - Reference data lookups

- Quality inspection reports
- In software
- Using document templates for data entry

Common business rule types:

- **Validity**: Implemented through rules such as:
  - **Format compliance:** Values have a pattern e.g., phone numbers
  - **Value domain membership:** Exists in a defined data domain. (Master and reference data)
  - **Range conformance:** Within a defined range of values
  - **Mapping conformance:** Maps to a domain (Reference data) e.g. State code to full name
- **Completeness or Value presence and record completeness:** Rules defining the conditions under which missing values are acceptable or unacceptable
- **Consistency rules:** Maintain a relationship between two data elements based on the actual values
- **Consistency may drive Definitional conformance rules:** Same understanding of data definitions across the organisation
- **Integrity rules:** To maintain a relationship between two or more data elements based on the actual values of those data elements
- **Timeliness validation:** Characteristics associated with expectations for accessibility and availability of data, and also when it was last updated
- **Uniqueness:** Rules that specify which entities must have unique representation and whether one and only one record exists for each real world object
- **Reasonableness**: Aggregation checks
  - Relative measure
  - Absolute measure
- **Accuracy verification:** Compare value to system of record or trusted source

1.3.4 Data Quality Improvement Lifecycle

Approach data quality improvement based on the technique of quality improvement in manufacture of physical products. A process is a series of steps that turn inputs into outputs. Outputs from one process become inputs to other processes and any step can be negatively affected and can impact data quality. Requirements must be defined along the whole data chain.

Use “plan-do-check-act” from problem solving model, the Shewhart/Deming Cycle.

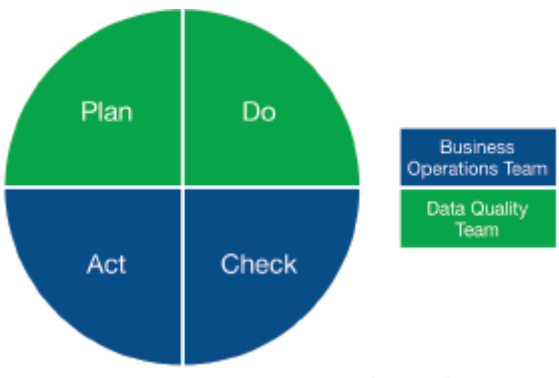


Figure 92 The Shewhart Chart with the role of a Data Quality Team indicated

- **Plan Stage**: DQ team assesses scope, impact and priority of known issues, and evaluates alternatives to address them at the root cause
- **Do stage**: DQ team leads efforts to address issues at the root cause and plans for ongoing monitoring of data
- **Check stage**: Business operations teams actively monitor data against standards. If data falls below, act to bring it to acceptable levels
- **Act stage**: Activities to address emerging data issues by business operations teams.

New cycles begin as:

- Existing measurements fall below thresholds
- New data sets come under investigation
- New DQ requirements emerge
- Business rules, standards, or expectations change

Dimension of Quality	Description
Plan (Data Quality Team)	1 Define High Quality Data 2 Define a Data Quality Policy 3 Define Scope of Initial Assessment 4 Perform Initial Data Quality Assessment 5 Identify & Prioritize Improvements 6 Define Goals for Data Quality Improvement
Do (Data Quality Team)	7 Develop and Deploy Data Quality Operations a. Develop Data Quality Operational Procedures

Dimension of Quality	Description
Check (Operations)	8. Develop and Deploy Data Quality Operations a. Correct Data Quality Defects b. Measure and Monitor Data Quality c. Report on Data Quality levels and findings
Act	Correct Data Quality Defects

Table 30 Data Quality Activities

1.3.5 Common Causes of Data Quality Issues

Data quality issues can arise at any point in the data lifecycle, and may have multiple causes: Data entry, data processing, system design and manual intervention in automated processes.

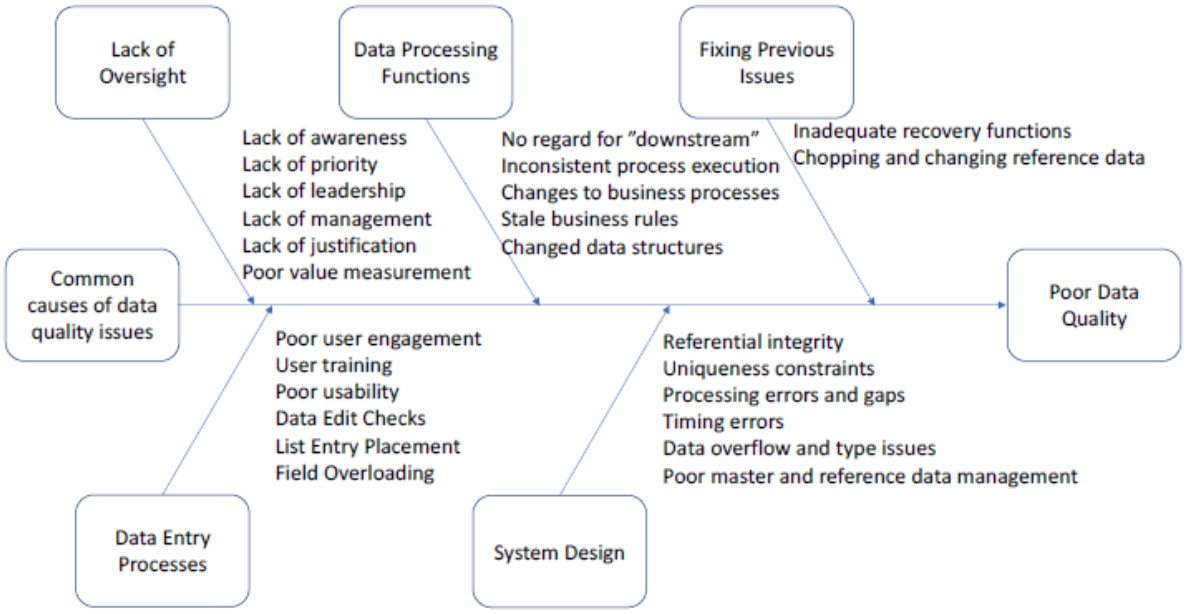


Figure 93 Sources of Data Quality Issues

1.3.5.1 Issues Caused by Lack of Oversight

Lack of organisational commitment to Data Quality (both management and governance). Barriers to effective management of data quality:

- Lack of awareness on the part of leadership and staff
- Lack of priority
- Lack of business governance
- Lack of leadership and management
- Difficulty in justification of improvements
- Inappropriate or ineffective instruments to measure value

1.3.5.2 Issues Caused by Data Entry Process

- **Poor user engagement:** If people creating the data see no value in entering the data they will take less care. E.g. expecting customers to enter too much
- **Training issues:** Lack of process knowledge, awareness on impact of incorrect data, or incentivised for speed rather than accuracy
- **Poor usability:** Anything that requires the user to work harder may lead to poor data through frustration
- **No data entry edit checks:** Have edits and controls to prevent input of incorrect data
- **Inadequate list entry placement**: Order of values in a dropdown list
- **Field overloading**: Reusing fields for different business purposes over time creates confusion

1.3.5.3 Issues Caused by Data Processing Functions

- **No regard for "downstream" processes**: in a siloed environment without understanding of where the data comes from or goes
- **Inconsistent business process execution**: Different user groups may use the data and processes differently. May be due to training or insufficient documentation
- **Changes to business processes**: New business rules and data quality requirements need to be incorporated into systems

- **Stale business rules:** Business rules may change over time, and should be periodically reviewed
- **Changed data structures:** Source system structure changes without informing downstream consumers or providing time for their changes

1.3.5.4 Issues Caused by System Design

- **Failure to enforce referential integrity**:
  - Duplicate data that breaks uniqueness rules
  - Orphan rows, in some reports and not others, leading to different values for the same calculation
  - Inability to upgrade due to restored or changed referential integrity requirements
  - Inaccurate data due to missing data being assigned default values
- **Failure to enforce uniqueness constraints:** Multiple copies of data instances in the same table result in overstated aggregation
- **Processing errors and gaps:** Data mapping or layout is incorrect or rules for processing incorrect data are inaccurate
- **Timing errors: S**ystems process data of a different currency level (up to date-ness)
- **Data overflows and type issues:** Assumptions in data model must be supported by actual data
- **Poor reference and Master Data Management:** Defects in these shared data sets can affect many business processes. Duplicates in the list of clients is a common challenge.

1.3.5.5 Issues Caused by Fixing Issues

Manual data patches directly to database. Most change the data in place, and can only be undone by a database restore.

2 Activities

2.1 Define a Data Quality Framework

Data Quality priorities must align with business strategy. Framework and methodology guides strategy and tactics and provides a means to measure progress and impacts. Methods to:

- Understand and prioritise business needs
- Identify data critical to meeting business needs
- Define business rules and DQ standards based on business requirements
- Assess data against expectations
- Share findings and get feedback from stakeholders
- Prioritise and manage issues
- Identify and prioritise opportunities for improvement
- Measure, monitor and report on data quality
- Manage Metadata produced through quality processes
- Integrate data quality controls into business and technical processes

2.2 Define High Quality Data

High quality data is fit for the purposes of data consumers. First focus on the data that, if it were of higher quality, would provide greater value to the organisation and its customers.

Start with Master Data, which is by definition, the most important to the organisation.

Understand business needs, define terms, identify pain points and start to find consensus about drivers and data quality priorities. Ask questions of stakeholders to determine the understanding of the business benefits of high-quality data and the impact of poor quality data

- What do stakeholders mean by "high quality data?"
- What is the impact of low quality data on business operations and strategy?
- How will higher quality data enable business strategy?
- What priorities drive the need for Data quality improvement?
- What is the tolerance for poor quality data?
- What governance is in place to support the Data Quality improvement?
- What additional governance structures will be needed?

Understand the current state of data quality by approaching the question from different perspectives:

- Understand business strategy and goals
- Pain points, risks and business drivers from stakeholders
- Direct assessment of data (profiling and analysis)
- Documentation of data dependencies in business processes
- Documentation of technical architecture and systems support for business processes

Prioritise opportunities based on benefits to the organisation.

2.3 Identify Dimensions and Supporting Business rules

- Business Rules
    - Describe expectations about the dimensions of Data Quality
    - Often not documented – need to reverse engineer through analysis of business processes, workflows, regulations, policies, standards, system edits, software code, triggers and procedures, and common sense
    - Business rules for how data is collected or created
    - Measurements to describe if data is fit for use in terms of DQ dimensions
    - Stakeholders give input and output requirements and pain points, what happens when data is missing or incorrect, how they recognise bad data

2.4 Perform an Initial Data Quality Assessment

Query the data to understand content and relationships. Data stewards, SMEs, data consumers and DQ analysts prioritise findings.

The goal of the initial assessment is to learn about the data to make an actionable plan for improvement:

- Define the goals to drive the work
- Identify data to be assessed. Start small
- Identify uses of the data and consumers of the data
- Identify known risks of the data
- Inspect data based on known and proposed rules
- Document levels of non-conformance and types of issues
- Perform in depth analysis based on initial findings to
    - Quantify findings
    - Prioritise issues based on business impact

    - Develop root cause hypotheses
- Confirm issues and priorities with data stewards, SMEs and data consumers
- Use findings as a foundation for planning
    - Remediation of issues, ideally at the root cause
    - Controls and processes to prevent issues from recurring
    - Ongoing controls and reporting

2.5 Identify and Prioritise Potential Improvements

Involves a combination of data analysis and discussion with stakeholders

- Requires identifying and prioritising potential improvements
- Profile large data sets to understand breadth of issues
- Interview stakeholders on issues
- Analyse business impact of issues
- Large-scale profiling efforts should focus on the most critical data

2.6 Define Goals for Data Quality Improvement

- Quick hits as well as long term strategic changes
- Address root causes
- Set achievable goals based on quantification of the business value of DQ improvements
- Determine ROI of fixes of issues based on:
    - Criticality of the data
    - Amount of affected data
    - Age of the data
    - Number and type of business processes impacted
    - No of stakeholders affected
    - Associated risks
    - Cost of root cause remediation
    - Cost of work arounds

2.7 Develop and Deploy Data Quality Operations

2.7.1 Manage Data Quality Rules

Define rules up front to:

- Set clear expectations
- Provide requirements for edits and controls to prevent data issues from being introduced
- Provide DQ requirements to vendors
- Foundation for DQ measuring

DQ rules should be managed as Metadata and should be:

- **Documented consistently:** Standards and templates for consistent format and meaning
- **Defined in terms of DQ Dimensions:** Help people understand what is being measured
- **Tied to business impact:** Connect standards and rules to their impact on organisational success
- **Backed by data analysis:** Test rules on actual data
- **Confirmed by SMEs**
- **Accessible to all data consumers**

2.7.2 Measure and Monitor Data Quality

Work of the DQ team focuses on measuring and reporting on quality. High level categories of DQ metrics:

- **Return on Investment:** Costs of improvement vs. benefits of improved DQ
- **Levels of quality:** Percentage of errors in a data set or across data sets
- **Data Quality trends:** Quality improvements over time
- **Data issue management metrics:**
  - Counts of issues by DQ dimensions
  - Issues per business function and statuses
  - Issues by priority and severity
  - Time to resolve issues

Two important reasons to implement operational Data Quality measurements:

1. **To inform consumers about levels of quality:**
   a. Based on data assessment and root cause analysis
   b. Critical data elements and relationships
2. **Manage risk that may be introduced by technical or business changes:**
   a. Knowledge of past problems to prevent risk of future problems
   b. Monitor functions similar to those that have had problems

In Figure 94, the Data Quality threshold is the dotted line in each graph. The Data Quality team progresses each of these trends:

- 2, 4, and 5 are ongoing issues (the team is monitoring progress to bring them back under threshold)
- 1, 3, 10 are today's issues (the team is mobilizing efforts to identify and address the trend)
- 7, 8, 11 are on today's watch list (the team is taking an interest but has not yet mobilized to address)
- 6 is a long-term intransigent issue (the team has tried a couple of ideas but is still looking for a permanent solution)
- 9, 12 are fine (the team are happy with these trends).

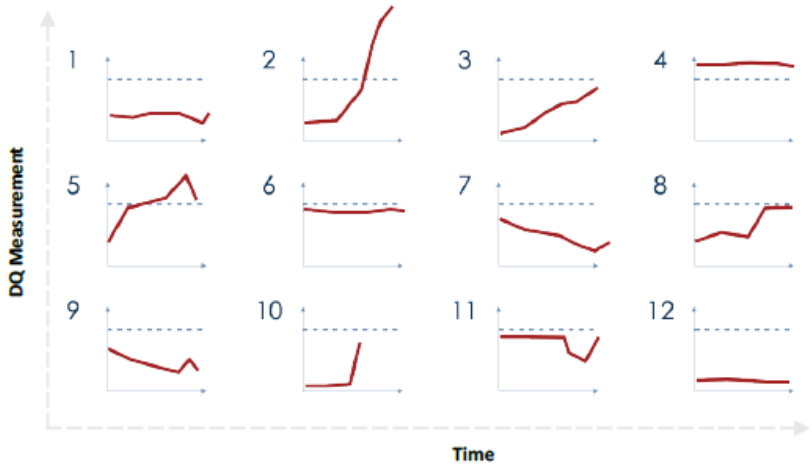


Figure 94 Illustrations of data defects over time

Measurements can be taken at three levels of granularity:

Table 31 Data Quality Monitoring Techniques

Granularity	In-stream (In-Process Flow) Treatment	Batch Treatment
Data Element	Edit checks in application Data element validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data Record	Edit checks in application Data record validation services Specially programmed applications	Direct queries Data profiling or analyzer tool
Data set	Inspection inserted between processing stages	Direct queries Data profiling or analyzer tool

Table 31 Data Quality Monitoring Techniques

2.7.3 Develop Procedures for Managing Data Issues

The key elements of an incident management process are:

- **Standardise Data Quality issues and activities**: Standard vocabulary to simplify classification and reporting
- **Provide an assessment process for data issues**:
- **Manage issue escalation procedures**: Specified in the DQ SLA. Incident management system implements escalation procedures
- **Manage Data Quality resolution workflow**: DQ SLA for monitoring, control, and resolution

The DQ Team must design and implement detailed operational procedures for:

- **Diagnosing issues:** Trace the lineage of the data to pinpoint potential **r**oot cause. Procedure describes:
  - Review issues in context of processing flows to isolate point where flaw is introduced
  - Evaluate whether there been environment changes that could have caused errors
  - Evaluate whether other process issues contributed
  - Determine whether there are issues with external data that could have affected this data
- **Formulating options for remediation:** Based on diagnosis, evaluate alternatives for addressing the issue:
  - Addressing non-technical such as lack of training, leadership support, accountability
  - Modification of systems to eliminate root causes
  - Developing controls to prevent issue
  - Introducing additional inspecting and monitoring
  - Directly correcting flawed data
  - Take no action based on cost and impact of correction versus the value of data correction
- **Resolving issues:** DQ team and business data owners determine the best way to solve issues:
  - Assess costs and merits of alternatives
  - Recommend planned alternatives
  - Provide a plan for developing and implementing the resolution
  - Implement the resolution

2.7.4 Establish Data Quality Service Level Agreements

- A DQ SLA specifies the organisation's expectation for response and remediation for DQ issues in each system. The SLA defines roles and responsibilities associated with performance DQ procedures
- SLA establishes time limits for notification generation, the names in the management chain and when escalation should occur
- SLA reporting can be scheduled or driven by business

2.7.5 Data Quality Response

Reporting should focus around:

- DQ Scorecard
- DQ trends
- SLA metrics
- DQ Issue management focussing on the status of issues and resolutions
- Conformance of the DQ team to governance policies
- Conformance of IT and business teams to DQ policies
- Positive effects of improvement projects

3 Tools

3.1 Data Profiling Tools:

- Produce high level statistics to identify patterns in the data
- Perform initial assessment of DQ
- First step in data analysis to identify potential issues
- DQ team members need to query more deeply to answer questions, find patterns that provide insights to root causes
- Enable assessment of large data sets

3.2 Business Rule Templates and Engines

- Capture expectation for data standard forms
- Templates facilitate communication between business and technical teams

3.3 Data Parsing and Formatting

Data Parsing is the process of analysing data using pre-determined rules to define its content or value:

- Valid and invalid values
- Matching specific patterns trigger actions

3.4 Data Transformation and Standardisation

Rule based transformations map data values in their original formats and patterns into target representation. Standardisation employs rules that capture context, linguistics and idioms recognised as common over time.

3.5 Data Enrichment

Data enhancement or enrichment is the process of adding attributes to data to increase its quality or usability.

- **Time/Date Stamp:** Root cause analysis to isolate the timeframe of the issue
- **Audit data:** Adds lineage for historical tracking and validation
- **Reference vocabularies:** Enhance understanding and control while bringing business context
- **Contextual information:** Adding context such as location, tagging
- **Geographic information:** Address standardisation and geocoding
- **Demographic information:** Customer data enhanced through demographic information
- **Psychographic information:** Segment target populations by behaviours
- **Valuation information:** Asset valuation, inventory and sale

3.6 Incident Management System

An incident tracking system should be used:

- Standardise data quality issues and activities
- Provide an assignment process for data issues
- Manage issue escalation procedures
- Manage data quality resolution workflow

4 Techniques

Standard sequence:

- Metrics – Measuring the quality of data
- Profiling – Finding statistical or relationship patterns
- Preventative or corrective actions – avoid DQ problems
- Root cause analysis – fundamentals of a problem
- Corrections – How to tidy up a data issue

4.1 Effective Metrics

Characteristics of informative metrics:

- **Measurability:** Quantifiable within a discrete range
- **Business relevance**: Correlate with the influence of the data on the key business expectations
- **Acceptability:** Data meets business expectation based on acceptability thresholds
- **Accountability/Stewardship**: Understood and approved by key stakeholders
- **Controllability:** Should trigger an action to improve data
- **Trending:** Measure DQ over time
- Express as a percentage where (r) is the rule being tested

$$ValidDQI(r) = \frac{(TestExecutions(r) - ExceptionsFound(r))}{TestExecutions(r)}$$

$$InvalidDQI(r) = \frac{(ExceptionsFound(r))}{TestExecutions(r)}$$

Example:

10000 tests of business rule (r) found 560 exceptions
Therefore: Valid DQ = (10000-560)/10000 = 94.4%
Invalid DQ = 560/10000 = 5.6%

Example of completeness

Business Rule	General	Example
Business Rule	Population of field is mandatory	Postal Code must be populated in the address table for 100%.
Measure	Count the number of records where data is populated, compare to the total number of records	Count populated: 700,000 Count not populated: 300,000 Total count: 1,000,000
Metrics (calculation)	Divide the obtained number of records where data is populated by the total number of records in the table or database and multiply it by 100 to get to percentage complete	Positive measure: 700,000/1,000,000*100 = 70% populated Negative measure: 300,000/1,000,000 *100 = 30% not populated
Status indicator	Acceptable if the business rule is met. Unacceptable is the business rule is not met.	Unacceptable. There are too much unpopulated Postal Codes.

Example of uniqueness

Business Rule	General	Example
Business Rule	There should be only one record per entity instance in a table	There should be one and only one current row per postal code on the Postal Codes master list: 0% duplicates.
Measure	Count the number of duplicate records identified; report on the percentage of records that represent duplicates	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000
Metrics (calculation)	Divide the number of duplicate records by the total number of records in the table or database and multiply it by 100	10,000/1,000,000*100 = 1.0% of postal codes are present on more than one current row
Status indicator	Acceptable if the business rule is met. Unacceptable is the business rule is not met.	Unacceptable. There are too many duplicates.

Example of timeliness

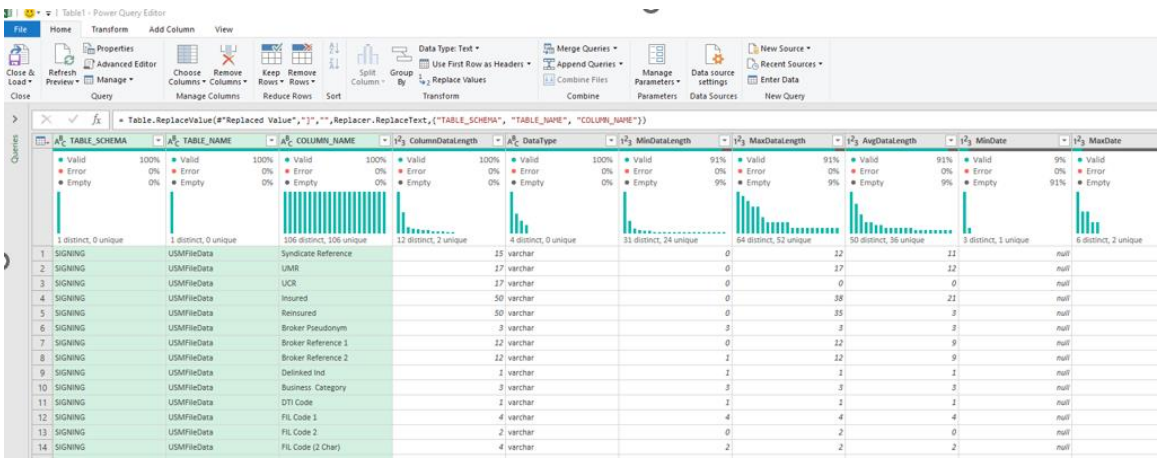
Business Rule	General	Example
Business Rule	Records must arrive within a scheduled timeframe	99% of equity market record should arrive within 5 minutes of being transacted
Measure	Count the number of records failing to arrive on time from a data service for business transactions to be completed	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000

Table 32 Data Quality Metric Examples

4.2 Data Profiling

Data profiling is a form of data analysis used to inspect the data and assess data quality. A profiling engine produces statistics that can be analysed to identify patterns in data content and structure.

- **Counts of NULLS:** Inspect to see if they are allowed
- **Max/Min value:** Identifies outliers
- **Max/Min Length:** Outliers in fields for specific length values
- **Frequency distribution** of values for individual columns
- **Data Type and Format:** Non-conformance or unexpected formats



4.3 Preventative Actions

- Sharable code modules that execute repeated DQ checks for developers
- Well-engineered code prevents many DQ problems

4.4 Effective Data Quality Metrics

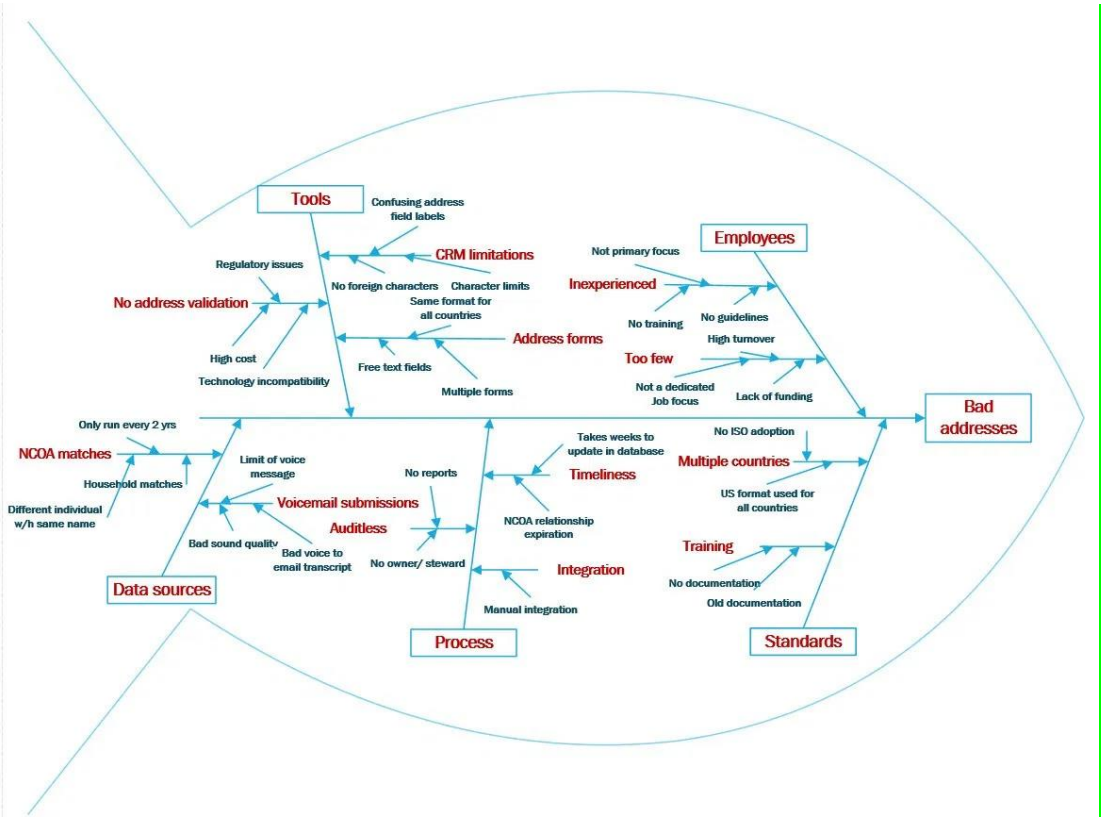
Ways to prevent poor quality data entering an organisation:

- **Establish data entry controls:** Data entry rules
- **Train data producers:** Value accuracy rather than speed
- **Define and enforce rules:** Create a 'data firewall', a table with all the business rules, to check quality before the data is used
- **Demand high quality data from data suppliers:** Examine processes to check structures, data sources and provenance
- **Implement data governance and stewardship:** Define roles and responsibilities
- **Institute formal change control:** Ensure changes are tested before implementing

4.5 Root cause Analysis

Common techniques:

- Pareto analysis (the 80/20 rule)
- Fishbone diagram analysis
- Track and trace
- Process analysis
- Five Whys (McGilvray, 2008)



The Ishikawa / Fish Bone Diagram (Author Unknown)

4.6 Corrective Actions

Perform data correction in three general ways:

- **Fully Automated:**
  - Rule based standardisation, normalisation and correction
  - No manual intervention
  - Requires environment with well-defined standards, rules and known error patterns
- **Manually-directed:**
  - Use automated tools with manual review before committing modified values to persistent storage
  - Environments where data sets require human oversight e.g., MDM
- **Manual:**
  - Best done through an interface with controls and edits which produces an audit trail
  - Avoid making manual corrections directly to the production environment

5 Implementation Guidelines

Improving DQ in an organisation requires changing how the people think and behave towards data. A DQ implementation needs to plan for:

- **Metrics on the value of data and the cost of poor data:** To raise organisational awareness and funding for improvements
- **Operating mode for IT/Business interactions:** Businesspeople know how important the data is and IT can translate definitions of data quality into queries that identify records which don't comply
- **Changes in how projects are executed:** Identify issues early and build data quality expectations into projects
- **Changes to business processes:** DQ team assesses processes and recommends changes
- **Funding for remediation and improvement projects:** Document costs and benefits of fixing data so that it can be prioritised
- **Funding for DQ operations:** Ongoing monitoring, reporting and fixing DQ issues

5.1 Readiness Assessment / Risk Assessment

Consider the following to assess the organisational readiness to accept DQ practices:

- **Management commitment to managing data as a strategic asset:** How much do they know about data as an asset, risks of poor-quality data, importance of data governance?
- **The organisation's current understanding of the quality of its data:** Pain points – helps identify and prioritise improvement projects
- **The actual state of the data:** Profiling, analysis and quantification of known pain points
- **Risks associated with data creation, processing or use:** Identify what can go wrong, and the potential damage to the organisation
- **Cultural and technical readiness for scalable data quality monitoring:** Requires a good collaborative relationship between IT and business teams

5.2 Organisation and Cultural Change

Employees need to think and act differently to produce and manage better quality data. Themes on how mindsets need to change:

- Lessening the grip of the Status Quo

- Dealing with politics
- Removing hero cultures

All employees raise DQ issues, ask for good quality data as consumers, and provide good quality data to others. Every person who touches the data can impact its quality.

6 Data Quality and the Other Knowledge Areas

All data management knowledge areas contribute to the quality of data. Cross-functional commitment and coordination produces high quality data.

6.1 Data Quality and Data Modelling and Design

DQ is enabled during the design phase of projects by allowing for data elements to:

- **Hold Data Quality values:** An extra data element in the design to mark a value as valid
- **Associate business rules:** Store rules in the data model e.g. Completeness

Modelling and ETL Tools impact DQ: Can improve quality if used with the data in mind, but can be detrimental if used without knowledge of the data.

6.2 Data Quality and Metadata Management

Metadata is created during the DQ process and must remain associated with the data. Metadata is the primary means of clarifying DQ expectations. If an address is marked valid, complete and reasonable by a process, this must be maintained as metadata with the address.

DQ and Metadata teams ensure that DQ requirements, rules, measurement results and documentation issues are stored in a metadata repository and shared across the organisation.

6.3 Data Quality and Master and Reference Data Management

- Master Data is critical data within DQ function
- Weak Master Data Management is a key source of poor Data Quality
  - Leads to choosing unreliable sources for data
  - Difficult to find the problem until the accuracy if the source is disproved
- DQ rule enforcement:
  - **Validity**: Domain of values managed as Reference Data
  - **Integrity**: On order must have an existing approved customer (which is Master Data)

6.4 Data Quality and Data Integration and Interoperability

Common source of errors within processes. Critical metadata must be maintained with the data. DQ can be monitored between processing stages at key integration points.

6.5 Data Quality and Data Governance

DQ is most effective when part of a Data Governance function. The DG Organisation can accelerate the work of the DQ function by:

- Setting priorities and providing direction
- Facilitating staff involvement and knowledge sharing
- Identifying and coordinating access to all involved in DQ decisions and activities
- Ensuring measurements of DQ are acted upon
- Sharing DQ inspection results, identifying opportunities for improvement, and consensus
- Resolving variations and conflicts

DQ function team works with a range of stakeholders and enables:

- Business process engineering
- Training staff to help teams implement process improvements
- Business and operational data stewards and data owners to:
    - Identify critical data
    - Define standards and quality expectations
    - Prioritise remediation of data issues

8 Data Quality Appendix

8.1 Data Quality ISO Standard

ISO 8000 is the international standard for data quality. ISO 8000 defines the characteristics that can be tested by any organisation in the data supply chain to objectively determine conformance.

ISO 8000 defines quality data as "portable data that meets stated requirements". Portable means that the data can be separated from a software application. Stated requirements must be clearly defined.

Describes the structure of DQ Management:

- DQ Planning
- DQ Control
- DQ Assurance
- DQ improvement

8.2 Further Reading on Data Quality Dimensions

DAMA UK white paper (2013) identified six core dimensions:

- **Completeness:** Proportion of data stored against 100%
- **Uniqueness:** No entity instance recorded more than once
- **Timeliness:** Degree to which data represents reality from the required point in time
- **Validity:** Data conforms to the syntax of its definition
- **Accuracy:** Degree to which data describes the real world
- **Consistency:** No difference found when comparing two or more representations of a thing to definitions

Others:

- **Currency:** Degree to which information is current / up to date
- **Precision:** Level of detail of the data element
- **Reasonableness:** Data pattern meets expectations

Other useful measurable characteristic (not listed as dimensions):

- **Usability:** Is the data understandable, simple, relevant, accessible, maintainable, and at the right level of precision?
- **Timing issues:** Is it stable yet responsive to legitimate change requests?
- **Flexibility:** Can it be repurposed? Is it easy to manipulate?
- **Confidence:** Are Data Governance processes in place? Is the data verified?
- **Value:** Good benefit case for the data? Is it being optimally used?

Relationship between Data Quality Dimensions and Data Quality Concepts:

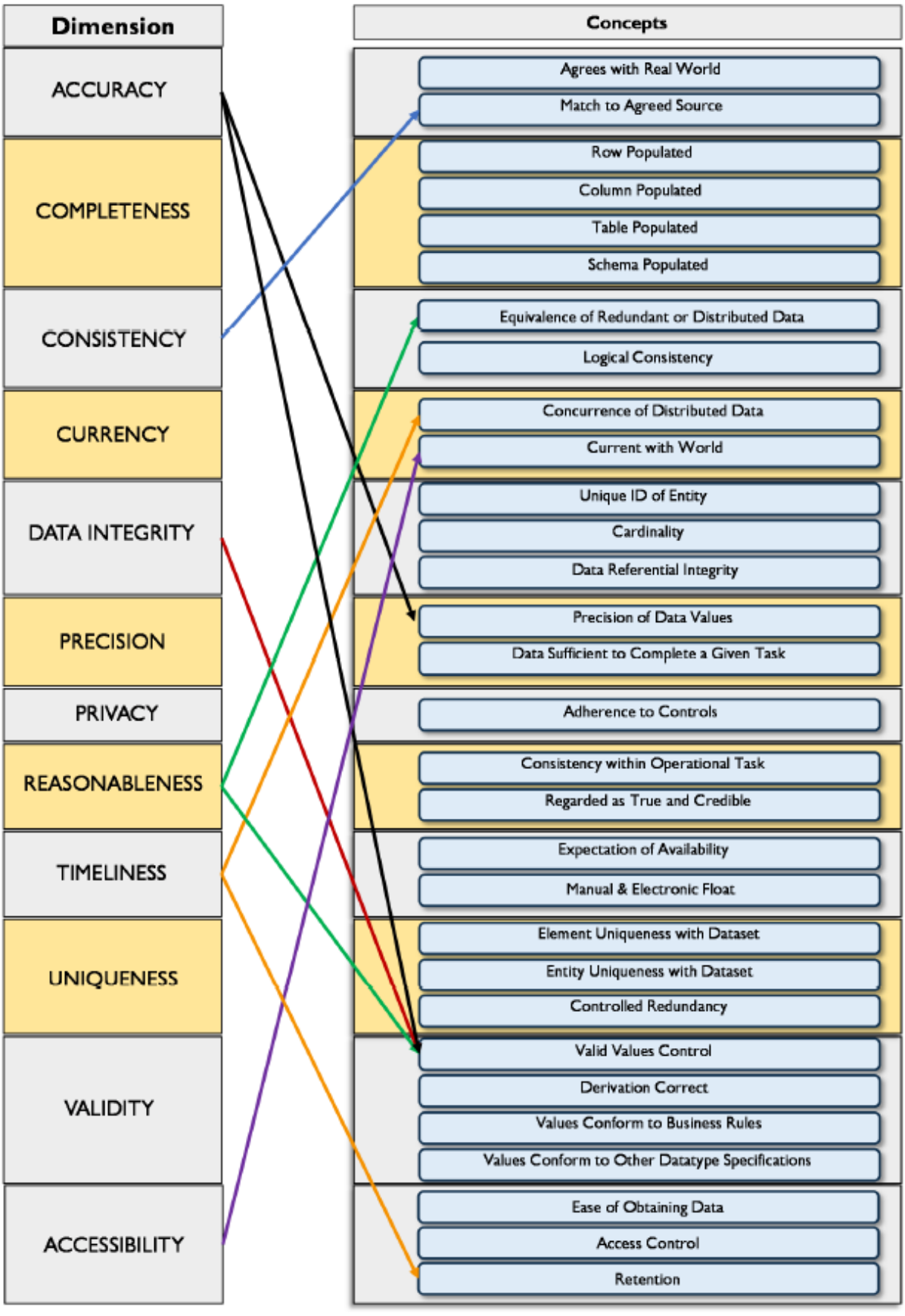
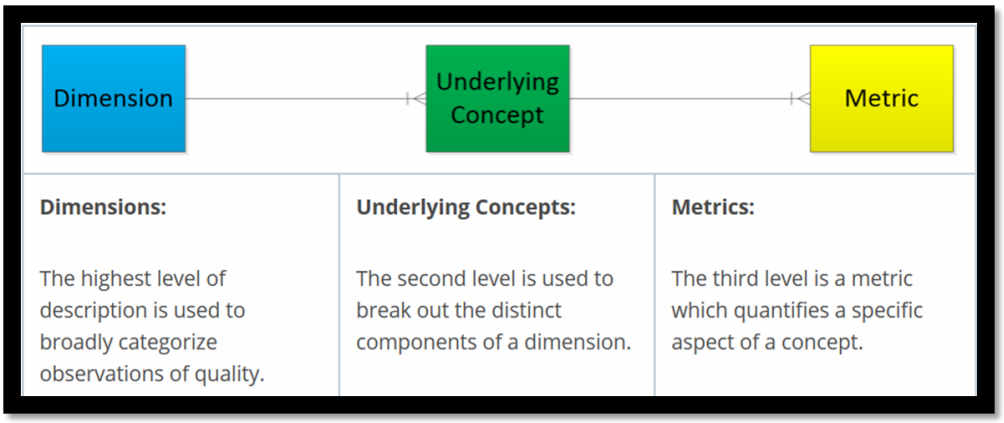


Figure 92 Relationship Between Data Quality Dimensions (Adapted from Myers 2013)


Dimension
Underlying Concept
Metric
Dimensions:
The highest level of description is used to broadly categorize observations of quality.
Underlying Concepts:
The second level is used to break out the distinct components of a dimension.
Metrics:
The third level is a metric which quantifies a specific aspect of a concept.

Relationship between Data Quality Dimensions, Underlying Concepts and Metrics (Author Unknown)

Validity	Consistent with the domain of values as well as other attribute values, Data conforms to the syntax of its definition
Completeness	All the required data is present (column, record or data set level)
Consistency	Data values in one data set are consistent with values in another data set
Integrity	Lack of incoherent values and broken relationships
Timeliness	The time the data needs to become accessible to a user after its capture
Currency	The date the data was last updated, and the likelihood it is still correct
Reasonableness	Data pattern meets expectations
Uniqueness/ Duplication	No entity instance recorded more than once within the data set
Accuracy	The degree that data correctly represents "real-life" entities
Precision	Level of detail of the data element

One line summary of all the DQ Dimensions

8.3 Statistical Process Control

SPC is a method to manage processes by analysing measurements of variation in process inputs, outputs or steps.

A process is a series of steps to turn inputs to outputs. SPC is based on the assumption that when a process with consistent inputs is executed consistently it will produce consistent outputs.

Uses measures of central tendency (mean, median, mode) and variability around a central value (range, variance, standard deviation) to establish tolerances for variances within a process.

Two types of variance:

- Common causes:
  - Inherent in the process
  - Process is in statistical control when only common causes, and range of variation (baseline) is established
- Special causes: Unpredictable or intermittent

SPC is used for control, detection and improvement. Early detection of unexpected variation simplifies root cause investigation.

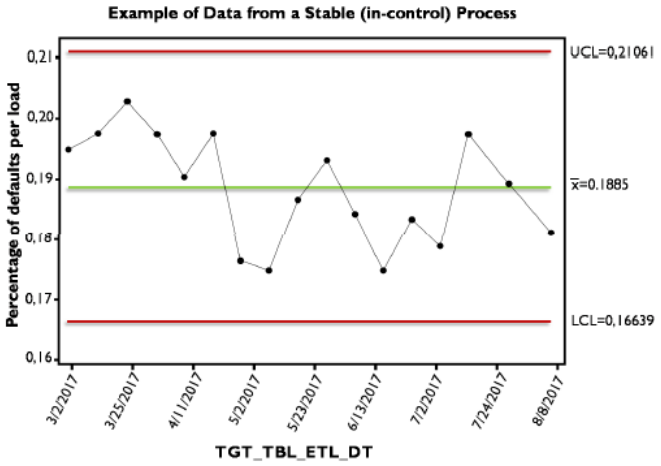


Figure 95 Control Chart of a Process in Statistical Control

Chapter 14
Big Data and Data Science

1 Introduction

Generate, store and analyse larger amounts of data:

- **Big Data:** Volume, variety and the velocity at which it is produced
- **Data Scientists:** People who mine and develop predictive, machine learning and prescriptive models and analytics
- **Data Science:** Applies data mining, statistical analysis and machine learning with data integration and data modelling to predict behaviours. Forward Looking. BI describes past trends – rear-window view

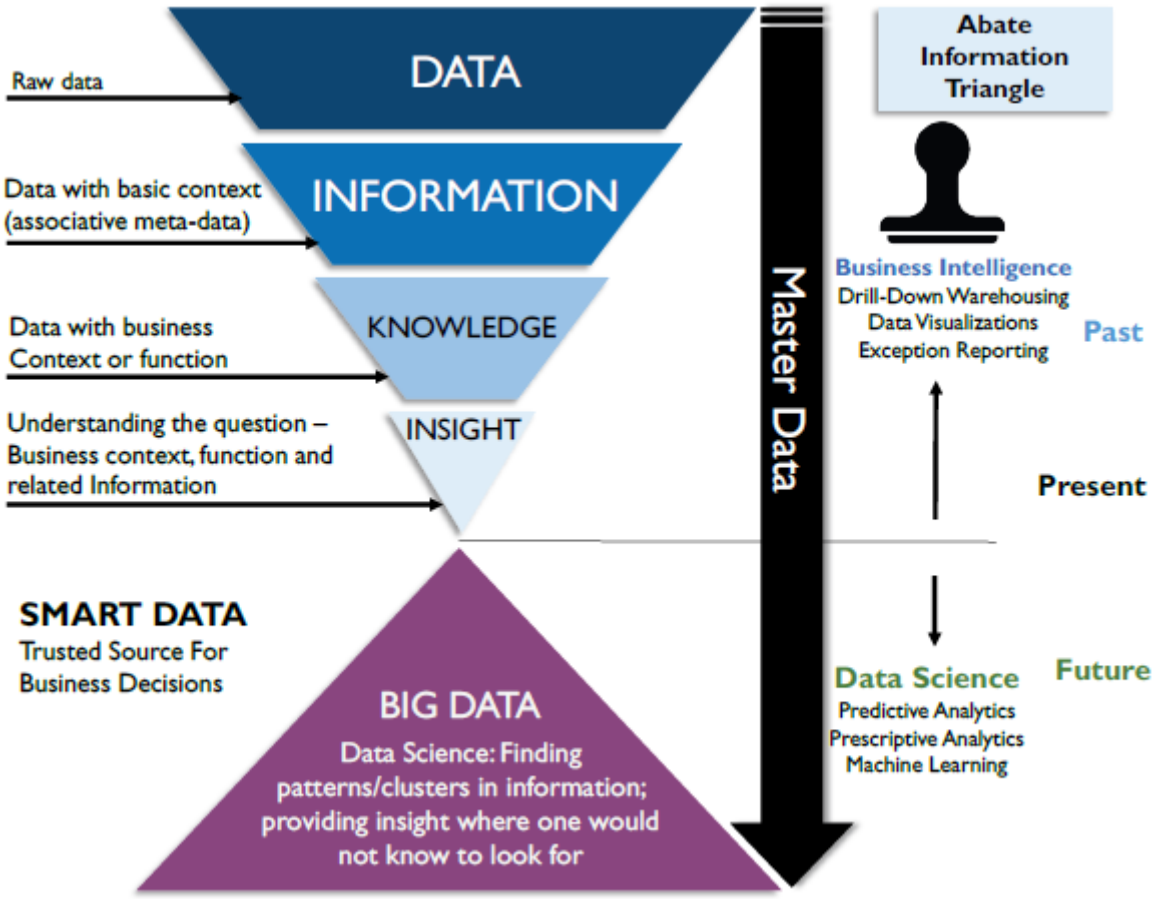


Figure 96 Abate Information Triangle

Different management and storage. Big Data relies on ELT – Loading and then transforming – and is not stored in the relational model. Speed and volume of data requires different approaches to Integration, Metadata and Data Quality management.

1.1 Business Drivers

- The desire to discover and act on business opportunities through applying techniques on diversely generated data

- Data Science can improve operations
- Machine learning for automation of complex time-consuming activities

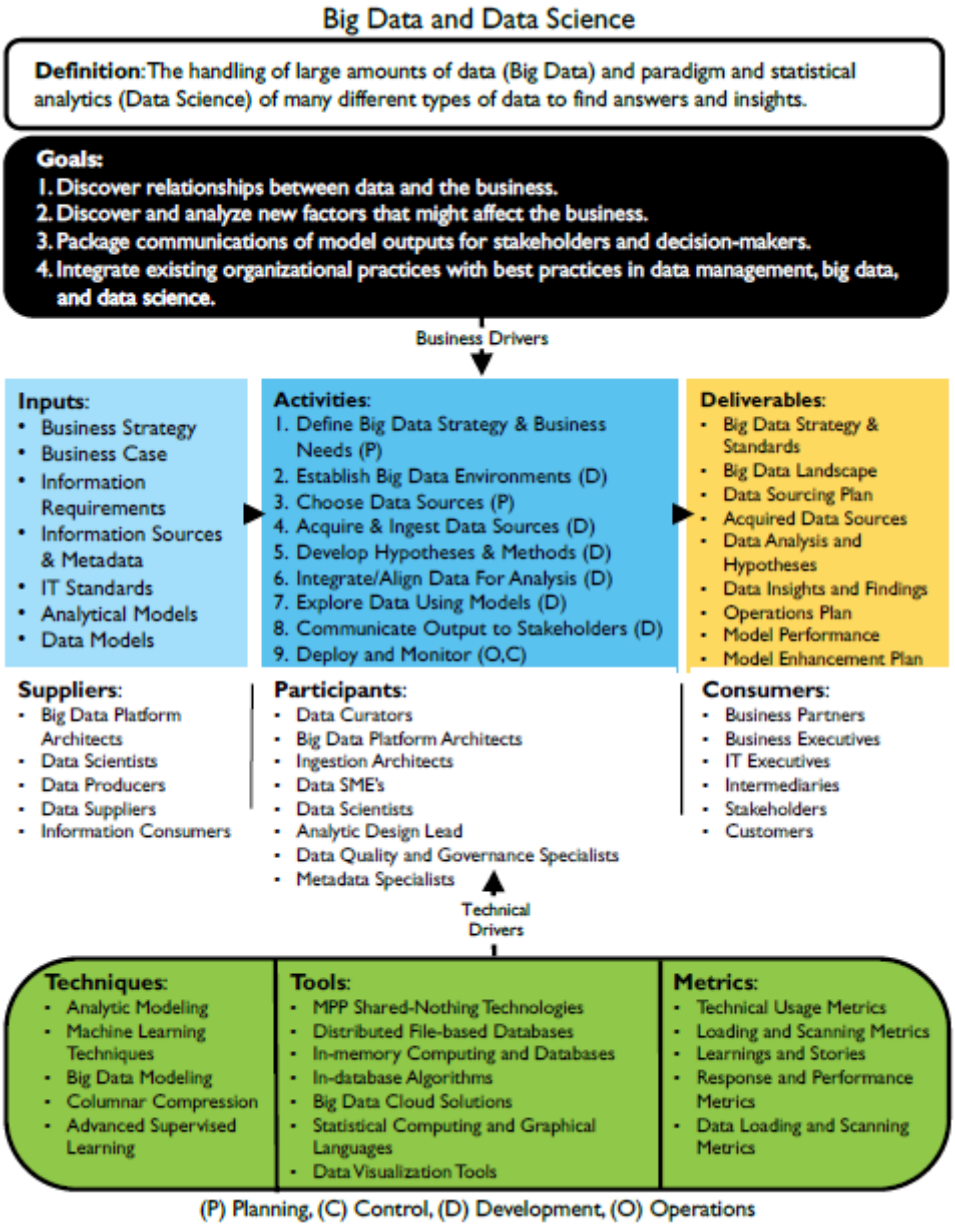

Big Data and Data Science
Definition: The handling of large amounts of data (Big Data) and paradigm and statistical analytics (Data Science) of many different types of data to find answers and insights.
Goals:
1. Discover relationships between data and the business.
2. Discover and analyze new factors that might affect the business.
3. Package communications of model outputs for stakeholders and decision-makers.
4. Integrate existing organizational practices with best practices in data management, big data, and data science.
Business Drivers
Inputs:
• Business Strategy
• Business Case
• Information Requirements
• Information Sources & Metadata
• IT Standards
• Analytical Models
• Data Models
Activities:
1. Define Big Data Strategy & Business Needs (P)
2. Establish Big Data Environments (D)
3. Choose Data Sources (P)
4. Acquire & Ingest Data Sources (D)
5. Develop Hypotheses & Methods (D)
6. Integrate/Align Data For Analysis (D)
7. Explore Data Using Models (D)
8. Communicate Output to Stakeholders (D)
9. Deploy and Monitor (O,C)
Deliverables:
• Big Data Strategy & Standards
• Big Data Landscape
• Data Sourcing Plan
• Acquired Data Sources
• Data Analysis and Hypotheses
• Data Insights and Findings
• Operations Plan
• Model Performance
• Model Enhancement Plan
Suppliers:
• Big Data Platform Architects
• Data Scientists
• Data Producers
• Data Suppliers
• Information Consumers
Participants:
• Data Curators
• Big Data Platform Architects
• Ingestion Architects
• Data SME's
• Data Scientists
• Analytic Design Lead
• Data Quality and Governance Specialists
• Metadata Specialists
Consumers:
• Business Partners
• Business Executives
• IT Executives
• Intermediaries
• Stakeholders
• Customers
Technical Drivers
Techniques:
• Analytic Modeling
• Machine Learning Techniques
• Big Data Modeling
• Columnar Compression
• Advanced Supervised Learning
Tools:
• MPP Shared-Nothing Technologies
• Distributed File-based Databases
• In-memory Computing and Databases
• In-database Algorithms
• Big Data Cloud Solutions
• Statistical Computing and Graphical Languages
• Data Visualization Tools
Metrics:
• Technical Usage Metrics
• Loading and Scanning Metrics
• Learnings and Stories
• Response and Performance Metrics
• Data Loading and Scanning Metrics
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 97 Context Diagram: Big Data and Data Science

1.2 Principles

- Big data management requires more discipline than relational data management
- Important to manage Metadata related to Big Data sources for accurate inventory of data sources, their origins and value.

1.3 Essential Concepts

1.3.1 Data Science

Developing predictive models that explore data content patterns. Based on hypothesis which may be statistically confirmed by historical data. Iterative inclusion of data sources into models that develop insights.

Data science depends on:

- **Rich data sources:** Potential to show otherwise invisible patterns in organisational or customer behaviour
- **Information alignment and analysis:** Techniques to understand data content and combine data sets for meaningful patterns
- **Information delivery:** Visualisations of results of insight gained from mathematical models
- **Presentation of findings and data insights:** Presentation and sharing of insights

DW / Traditional BI	Data Science	
Descriptive	Predictive	Prescriptive
Hindsight	Insight	Foresight
Based on history: What happened? Why did it happen?	Based on predictive models: What is likely to happen?	Based on scenarios: What should we do to make things happen?

Table 33 Analytics Progression

1.3.2 The Data Science Process

Data Science follows the scientific method of refining knowledge by making observations and testing hypotheses, observing results and formulating theories to explain results. The output of each phase in the diagram is input to the next.

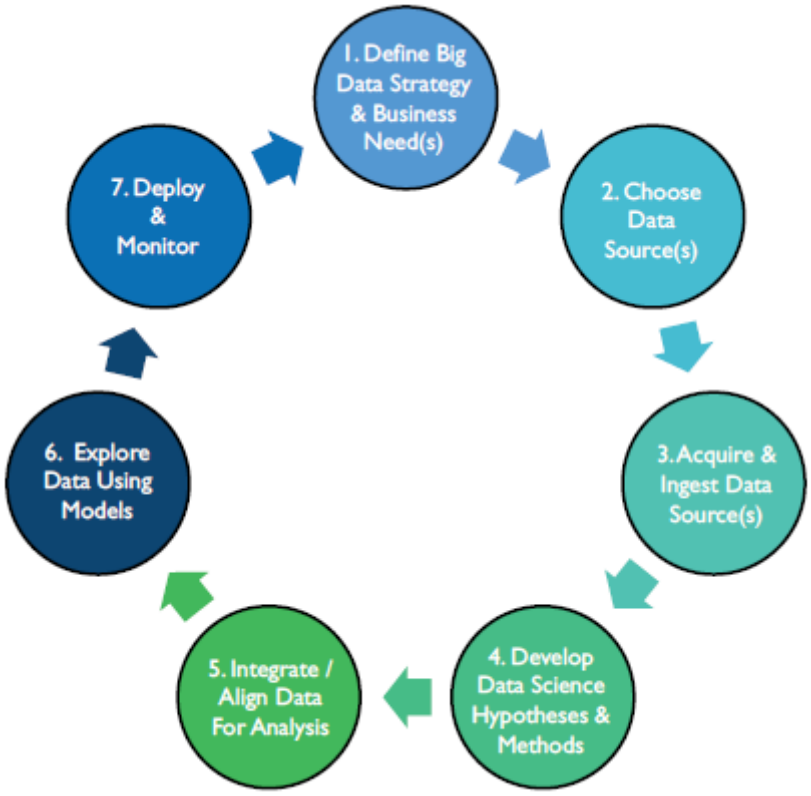

1. Define Big Data Strategy & Business Need(s)
2. Choose Data Source(s)
3. Acquire & Ingest Data Source(s)
4. Develop Data Science Hypotheses & Methods
5. Integrate / Align Data For Analysis
6. Explore Data Using Models
7. Deploy & Monitor

Figure 98 Data Science Process

- **Define Big Data strategy and business needs:** Requirements that identify desired outcomes with measurable benefits
- **Choose data sources:** Identify gaps in current data base and find sources

- **Acquire and ingest data stores**
- **Develop Data Science hypotheses and methods**
- **Integrate and align data for analysis:** Data integration and cleansing for quality (Wrangling and Munging)
- **Explore data using models:**
  - Apply statistical analysis and machine learning algorithms
  - Validate, train and evolve model
  - New hypotheses may be introduced
- **Deploy and monitor:** Often becomes a warehousing project

1.3.3 Big Data

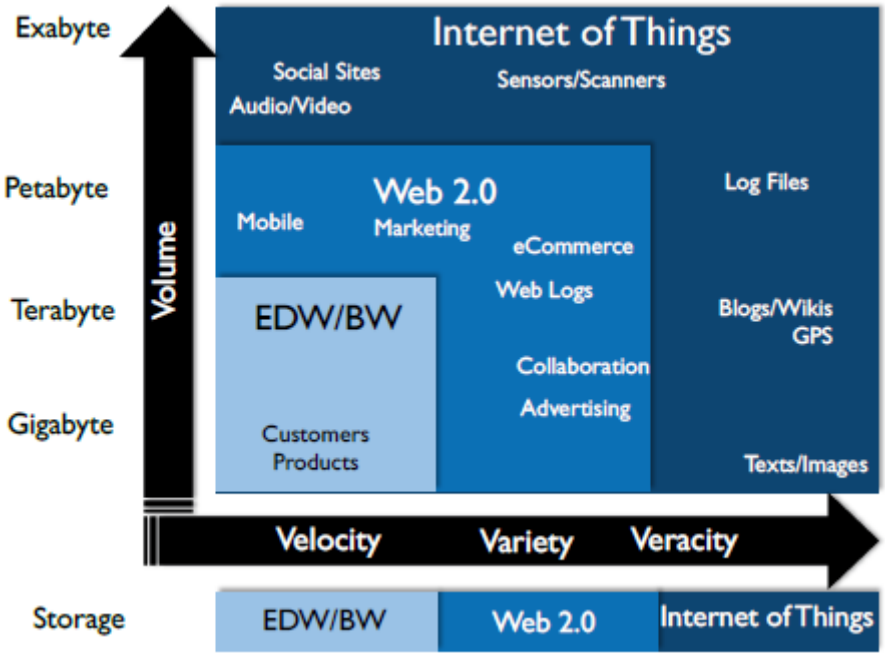


Figure 99 Data Storage Challenges

- **Volume:** Amount – billions of records - >100 Terabytes
- **Velocity:** Speed at which data is captured, generated or shared – often analysed at real-time
- **Variety / Variability:** Forms data is captured or delivered
- **Viscosity:** How difficult the data is to use and integrate
- **Volatility:** How often the data changes
- **Veracity:** How trustworthy the data is

1.3.4 Big Data Architecture Components

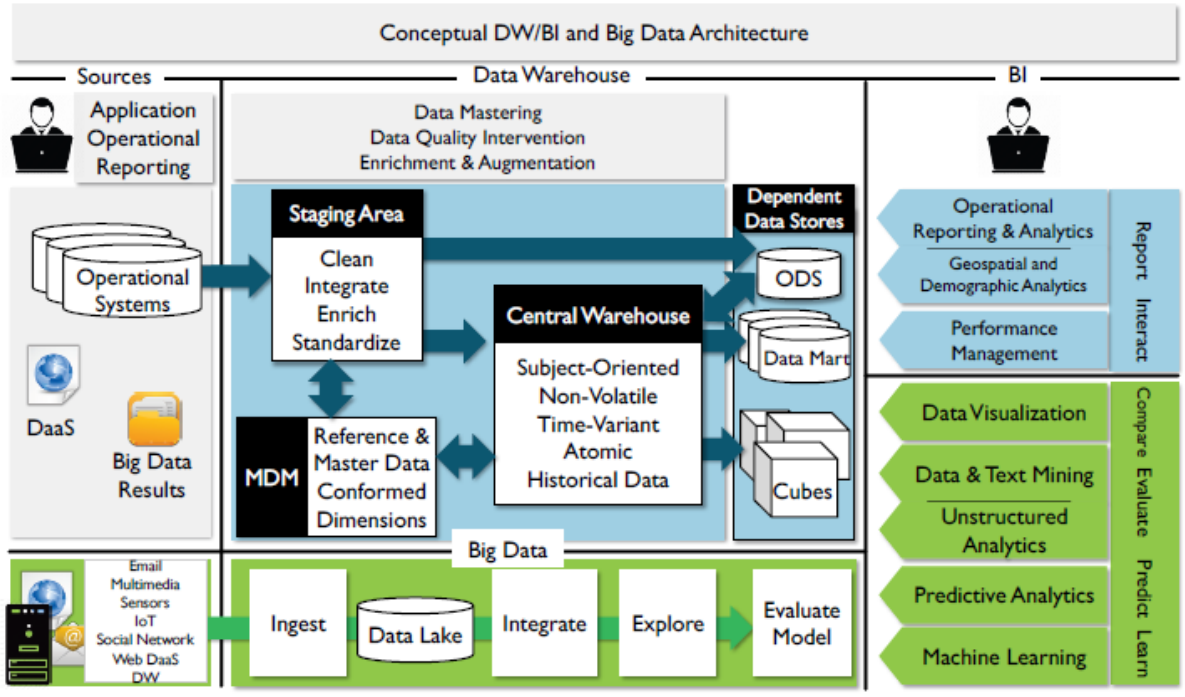


Figure 100 Conceptual DW / BI and Big Data Architecture

1.3.5 Sources of Big Data

Internet of Things, email, social media, online orders, video games, phones, POS devices, surveillance systems, sensors, medical monitoring, satellites, military etc.

1.3.6 Data Lake

An environment where a vast amount of data of various types and structures can be ingested, stored, assessed and analysed:

- Environment for Data Scientists to mine and analyse data
- Central storage for raw data with minimal transformation
- Alternate storage for detailed historical data warehouse data
- Online archive for records
- Environment to ingest streaming data with automated pattern identification

Manage Metadata to prevent it becoming a data swamp.

1.3.7 Services-Based Architecture

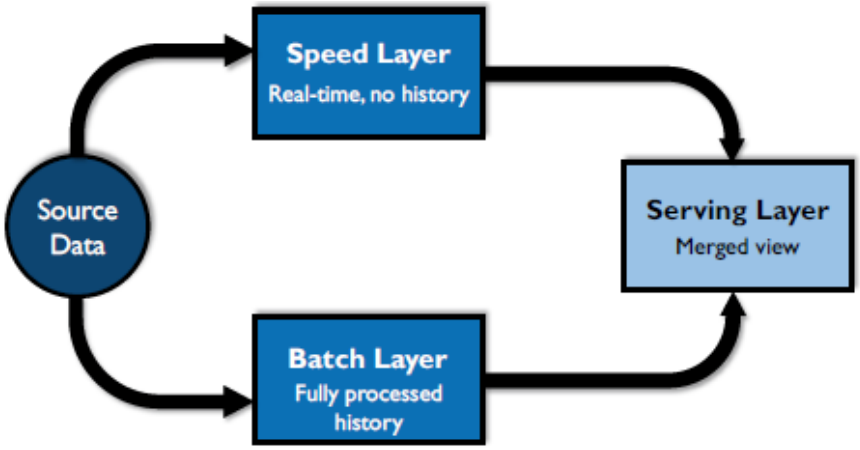


Figure 101 Services-based Architecture

Also called Lambda Architecture, referred to in Ch 5 p181, CAP / Brewers Theorem.

SBA provides immediate (maybe not complete or accurate) data. Updates complete, accurate historical data set using same source.

SBA has 3 main components:

- **Batch layer:** A data lake, historical, structure-over-time component, every transaction is an **insert**
- **Speed layer:** Only real-time data, Operational Data Store (ODS), all transactions are **updates** or **inserts**
- **Serving layer:** Interface to join batch and speed layers, uses Metadata to determine where to **serve** the data

Data is loaded into both Batch and Speed layers simultaneously, creating a current state, and a history layer.

1.3.8 Machine Learning

The construction and study of learning algorithms. Subfield of Artificial Intelligence. Three types:

- **Supervised learning:** Based on generalised rules e.g., separating SPAM emails
- **Unsupervised Learning:** Identifying hidden patterns i.e., data mining
- **Reinforcement learning:** Based on achieving a goal e.g., beating a chess opponent

Ethical implications:

- Transparency:
  - It is not clear how Deep Learning Neural Networks (DLNN) learn
  - Need to see how decisions are made

1.3.9 Sentiment Analysis

Looking for key words in semi-structured data using Natural Language Processing (NLP) to look for sentiment. Requires understanding of the meaning of a post.

1.3.10 Data and Text Mining

Data Mining is an offshoot of machine learning that reveals patterns in data using algorithms. Unsupervised learning where algorithms are applied to a data set without knowledge of outcomes, to reveal patterns and relationships.

Text mining analyses documents with text analysis and data mining techniques to classify content automatically into workflow guided and SME-directed ontologies.

Data and text mining techniques:

- **Profiling:** Characterise behaviour norms of an individual, group or population for anomaly detection (e.g., fraud)
- **Data reduction:** Make a similar smaller data set from a large one for easier analysis
- **Association:** Unsupervised learning process to find relationships based on transactions
- **Clustering:** Group elements by shared characteristics
- **Self-organising maps:** Neural network method of cluster analysis by reducing dimensionality in the evaluation space. Also called Kohonen Maps or topologically ordered maps

1.3.11 Predictive Analytics

Subfield of supervised learning. The development of probability models based on variables related to possible events. When it receives some information, the model triggers a response by the organisation. A forecast is a simple predictive model.

1.3.12 Prescriptive Analytics

Prescriptive analytics anticipates what will happen, when it will happen and implies why it will happen. Shows implications of various decisions, and can suggest how to take advantage of an opportunity or avoid a risk.

1.3.13 Unstructured Data Analytics

An iterative process of scanning and tagging to add hooks to unstructured data, to allow filtering and linking to related structured data.

1.3.14 Operational Analytics

BI or streaming analytics.

1.3.15 Data Visualisation

The process of interpreting concepts, ideas and facts by using pictures or graphical representations.

1.3.16 Data Mashups

Combine data and internet-based services to create visualisation for insight and analysis.

2 Activities

2.1 Define Big Data Strategy and Business Needs

Must be aligned with overall business strategy:

- What problems does the organisation need analytics to solve?
- What data sources to use or acquire?
- The timeliness and scope of the data to provision
- The impact on and relation to other data structures
- Influences to existing modelled data

2.2 Choose Data Sources

Big Data comes from many internal and external sources. Evaluate the Quality and reliability, and know its origin (provenance), format, what elements represent, how it connects to other data, and how frequently it will be updated.

Review available data sources, processes that create the data, and manage the plan for new sources.

- **Foundational data:** e.g., POS in sales analysis
- **Granularity:** Most granular form is ideal
- **Consistency:** Select data that appears appropriately and consistently across visualisations
- **Reliability:** Use trusted, authoritative sources
- **Inspect / profile new sources:** Test changes before adding new data sets

Risks:

- Privacy
- Filters which may introduce bias

2.3 Acquire and Ingest Data Sources

Capture critical Metadata about the source (origin, size, currency and additional knowledge about the content) when ingesting data sources into the Big Data environment. Ingestion engines may profile the data. Provides information on how to integrate with other data sets, such as Master Data or historical warehouse data, also how to train and validate models.

2.4 Develop Data Hypotheses and Methods

Data science entails building statistical models that find correlations and trends within and between data sets to find insights within the data. Models should be tested for a range of outcomes. Models depend on the quality of input data.

2.5 Integrate / Align Data for Analysis

Preparing data for analysis involves understanding what is in the data and links between data from various sources.

- Daily data would have to be aggregated to link to monthly data
- Common key
- Similarity and record linking algorithms
- Clustering used to determine groupings

2.6 Explore Data Using Models

- **Populate the Predictive Model**
- **Train the Model:** Repeated runs of the model against the data resulting in changes to the model
- **Evaluate the Model:**
    - Evaluated and validated against training sets
    - Ethical training to remove the biases of the creators
- **Create Data Visualisations:** Ensure the visualisation addresses the audience

2.7 Deploy and Monitor

- Expose Insights and Findings
- Integrate with additional Data Sources

3 Tools

3.1 MPP (Massively Parallel Processing) Shared-nothing Technologies and Architecture

A system that automatically distributes data and parallelises query workload across all available hardware.

Data is partitioned across multiple processing nodes each processing data locally. Communication is controlled by a master host and occurs over a network. There is no disk sharing or memory contention. Linearly scalable. Distribution of workload to processor level. Speeds up analytical functions.

3.2 Distributed File-based Databases:

- Open source Hadoop stores files of any type
- Language used is called MapReduce:
  - **Map:** Identify and obtain data to be analysed
  - **Shuffle:** Combine data according to desired analytical patterns
  - **Reduce:** Remove duplication and perform aggregation to reduce the size of the data set to only what is required

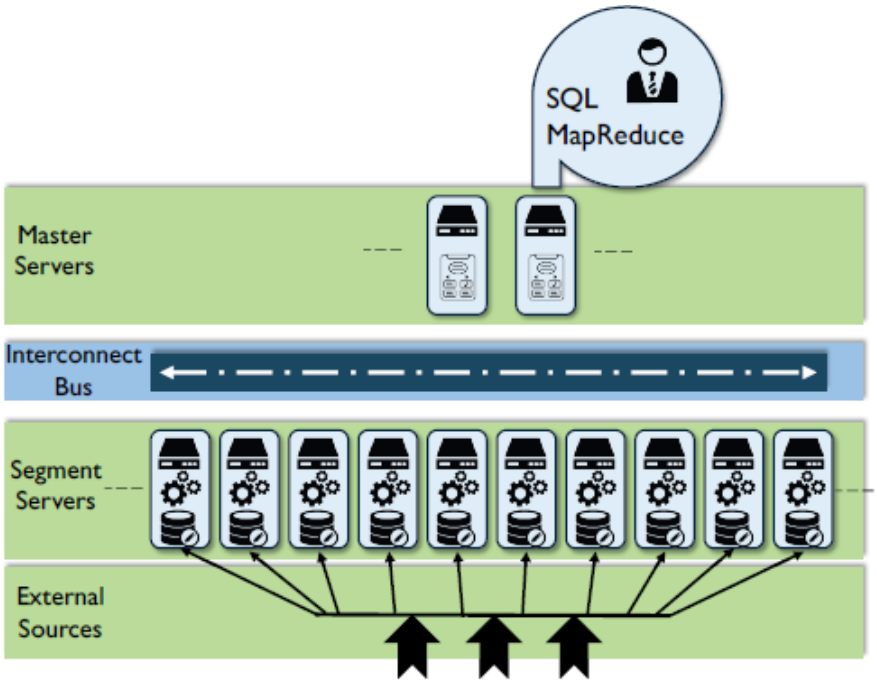


Figure 102 Columnar Appliance Architecture

3.3 In-database Algorithms

Each processor in an MPP Shared-nothing platform can run queries independently. Computation close to data reduces time for complex algorithms.

3.4 Big Data Cloud Solutions

Vendors provide cloud storage and enhancements.

3.5 Statistical computing and graphical languages

R – scripting language with statistical computing and graphics.

3.6 Data Visualisation Tools

- **Traditional visualisation tools:** both data and graphical
- **Information graphics / Infographics:** Insights, changes in data over time. Interactive, sophisticated analysis, adherence to visualisation best practices

4 Techniques

- **Analytic Modelling:** Different depths of analysis:
  - **Descriptive modelling:** Summarises data structures in a compact manner
  - **Explanatory modelling:** statistical models for testing a hypothesis
- **Big Data Modelling:** Data needs to be integrated, specified and managed by applying traditional Enterprise Architecture principles

5 Implementation Guidelines

5.1 Strategy Alignment

Strategically aligned with business objectives. Documents goals, approach and governance principles. Strategy deliverables should account for managing:

- Information lifecycle
- Metadata
- Data Quality
- Data acquisition
- Data access and security
- Data governance
- Data privacy
- Learning and adoption
- Operations

5.2 Readiness Assessment / Risk Assessment

Critical success factors:

- **Business relevance:** Big Data/Data Science initiatives must enforce business function
- **Business readiness:**
  - Prepared for long term delivery?
  - Committed to establishing centres of excellence to sustain the product?
  - How broad is the skill gap?
- **Economic viability:** Assessment of ownership costs – buying vs leasing
- **Prototype:** Can the solution be prototyped?

5.3 Organisation and Cultural Change

Need a communications program to engage stakeholders and a centre of excellence to provide training, best practices, knowledge management and communication across developer, designer, analyst and data consumer communities.

Cross functional roles:

- **Big Data platform architect:** Hardware, OS, file systems and services
- **Ingestion architect:** Data analysis, systems of record, modelling

- **Metadata specialist:** Metadata interfaces, architecture and contents
- **Analytic Design Lead:** End user analytic design
- **Data Scientist:** Architectural and model design based on statistical knowledge

6 Big Data and Data Science Governance

The enterprise view of data should drive decisions on sourcing, sharing, Metadata, enrichment and access.

6.1 Visualisation Channels Management

Alignment of the appropriate visualisation tools at the right level of complexity for the user community.

6.2 Data Science and Visualisation Standards

Vital for customer-facing and regulatory-facing content:

- Tools standards by analytic paradigm, user community, subject area
- Requests for new data
- Data set process standard
- Processes to avoid biased results:
    - Data inclusion and exclusion
    - Assumptions in the models
    - Statistical validity of results
    - Validity of interpretation of results
    - Appropriate methods applied

6.3 Data Security

- Agree upon levels of access for authorised personnel
- Mask data for those not authorised
- Use encryption for highly sensitive data
- Recombination measures the ability to reconstitute sensitive or private data and must be managed
- Outcomes of analysis may violate privacy

6.4 Metadata

Managed as part of data ingestion else data lake becomes a swamp.

6.5 Data Quality

Data Quality is a measure of deviation from expected result, the smaller the difference the higher the quality. Mature Big Data organisations scan data inputs using data quality tools to understand the information within:

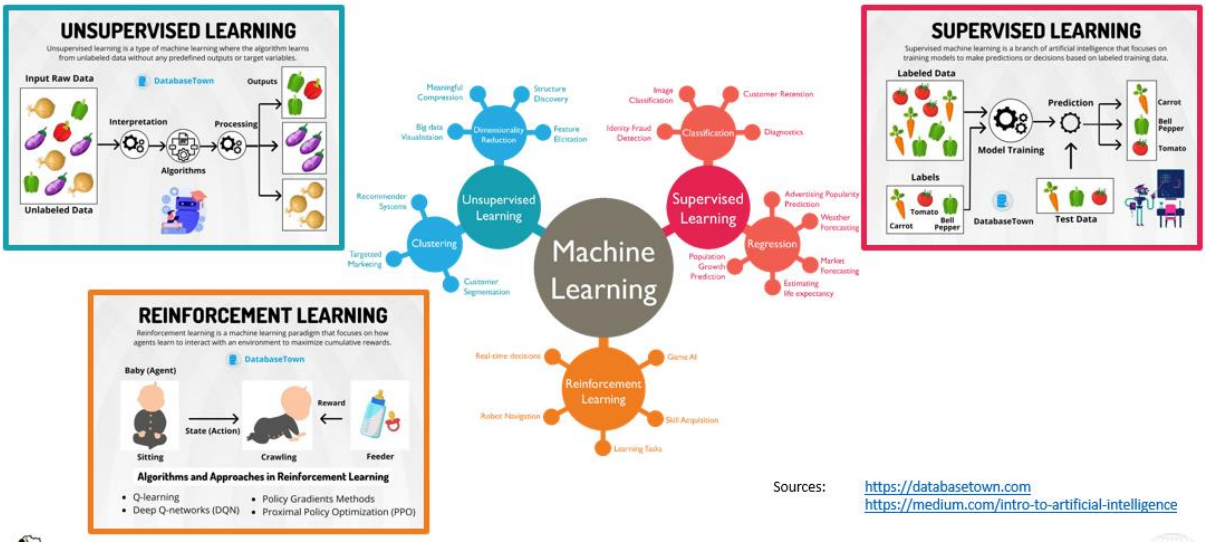
- **Discovery:** Where information resided in the data set
- **Classification:** What type of information based upon standardised patterns
- **Profiling:** How data is populated and structured
- **Mapping:** What other data sets can be mapped to these values

6.6 Metrics

- **Technical Usage Metrics:**
    - Look for data hot spots to manage distribution and performance

- Growth rates for capacity planning

- **Loading and scanning metrics:**
  - Ingestion rate
  - Interaction with the community
  - Provided by execution logs
- **Learnings and stories:**
  - Counts and accuracy of models and patterns
  - Revenue realised for identified opportunities
  - Cost reduction from avoiding identified threats



Sources: https://databasetown.com
https://medium.com/intro-to-artificial-intelligence

Chapter 15
Data Management Maturity Assessment

1 Introduction

Capability Maturity Assessment (CMA) is an approach to process improvement based on the Capability Maturity Model (CMM) that describes how processes evolve from ad hoc to optimal.

Progression happens in a set order:

- **Level 0:** Absence of capability. No formal process for managing data. Very few organisations exist at Level 0
- **Level 1:** Initial or ad hoc: success depends on the capability of individuals
- **Level 2:** Repeatable: Minimum process discipline is in place
- **Level 3:** Defined: Standards are set and used
- **Level 4:** Managed: Processes are quantified and controlled
- **Level 5:** Optimised: Process improvement goals are quantified

Criteria are described across process features related to execution, level of automation, policies, controls and/or process details.

A Data Management Maturity Assessment (DMMA) evaluates DM overall, within a knowledge area of even a process. Used as a bridge between business and IT.

DMMA provides a common vocabulary and a stage-based roadmap to improvement across knowledge areas.

1.1 Business Drivers

Reasons organisations conduct capability maturity assessments:

- **Regulation:** Minimum level of maturity required
- **Data Governance:** The data Governance function requires a maturity assessment for planning and compliance
- **Organisational readiness for process improvement:** Begin by assessing current state
- **Organisational change:** e.g., Challenge posed by a merger. DMMA provides input
- **New technology:** To understand the likelihood of success in adopting new technology
- **Data management issues:** Baseline current state in order to make better decisions on change

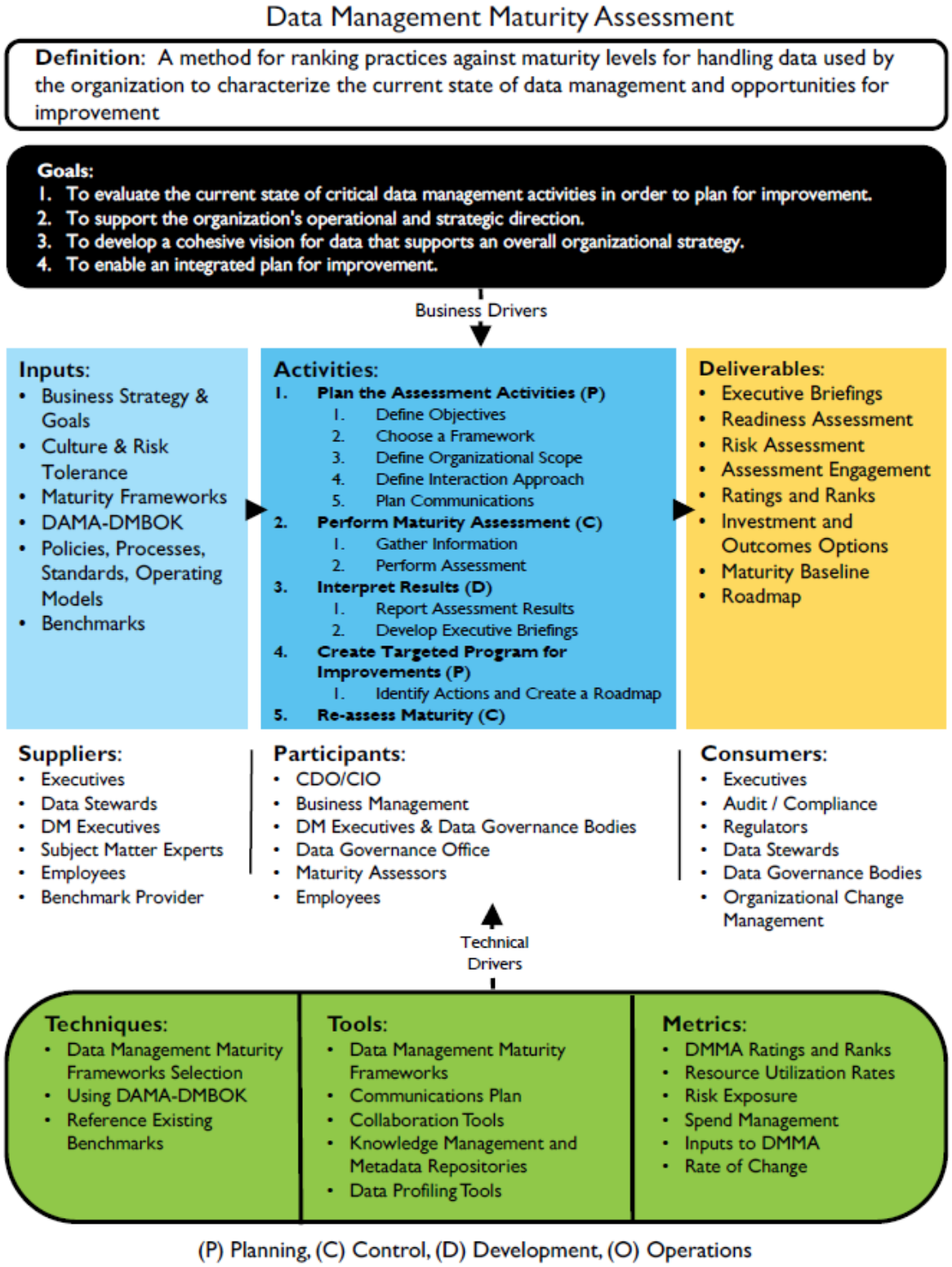

Data Management Maturity Assessment
Definition: A method for ranking practices against maturity levels for handling data used by the organization to characterize the current state of data management and opportunities for improvement
Goals:
1. To evaluate the current state of critical data management activities in order to plan for improvement.
2. To support the organization's operational and strategic direction.
3. To develop a cohesive vision for data that supports an overall organizational strategy.
4. To enable an integrated plan for improvement.
Business Drivers
Inputs:
• Business Strategy & Goals
• Culture & Risk Tolerance
• Maturity Frameworks
• DAMA-DMBOK
• Policies, Processes, Standards, Operating Models
• Benchmarks
Activities:
1. Plan the Assessment Activities (P)
1. Define Objectives
2. Choose a Framework
3. Define Organizational Scope
4. Define Interaction Approach
5. Plan Communications
2. Perform Maturity Assessment (C)
1. Gather Information
2. Perform Assessment
3. Interpret Results (D)
1. Report Assessment Results
2. Develop Executive Briefings
4. Create Targeted Program for Improvements (P)
1. Identify Actions and Create a Roadmap
5. Re-assess Maturity (C)
Deliverables:
• Executive Briefings
• Readiness Assessment
• Risk Assessment
• Assessment Engagement
• Ratings and Ranks
• Investment and Outcomes Options
• Maturity Baseline
• Roadmap
Suppliers:
• Executives
• Data Stewards
• DM Executives
• Subject Matter Experts
• Employees
• Benchmark Provider
Participants:
• CDO/CIO
• Business Management
• DM Executives & Data Governance Bodies
• Data Governance Office
• Maturity Assessors
• Employees
Consumers:
• Executives
• Audit / Compliance
• Regulators
• Data Stewards
• Data Governance Bodies
• Organizational Change Management
Technical Drivers
Techniques:
• Data Management Maturity Frameworks Selection
• Using DAMA-DMBOK
• Reference Existing Benchmarks
Tools:
• Data Management Maturity Frameworks
• Communications Plan
• Collaboration Tools
• Knowledge Management and Metadata Repositories
• Data Profiling Tools
Metrics:
• DMMA Ratings and Ranks
• Resource Utilization Rates
• Risk Exposure
• Spend Management
• Inputs to DMMA
• Rate of Change
(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 103 Context Diagram: Data Management Maturity Assessment

1.2 Goals and Principles

Goals:
1. To evaluate the current state of critical data management activities in order to plan for improvement.
2. To support the organization's operational and strategic direction.
3. To develop a cohesive vision for data that supports an overall organizational strategy.
4. To enable an integrated plan for improvement.

Primary goal is to evaluate the current state of maturity to plan for improvement. The organisation is placed on the maturity scale by clarifying specific strengths and weaknesses.

Benefits:

- Educates stakeholders about data management concepts, principles and practices
- Clarify stakeholder roles and responsibilities in relation to organisational data
- Highlight the need to manage data as a critical asset
- Broaden recognition of data management activities across the organisation
- Improve the collaboration necessary for effective data governance

1.3 Essential Concepts

1.3.1 Assessment Levels and Characteristics

CMMs define 5 or 6 levels of maturity:

- **Level 0: No Capability:** No formal data management practices or processes
- **Level 1 Initial/Ad Hoc:** Assessment criteria – Presence of any process controls
    - Limited tool set
    - Little or no governance
    - Reliant on a few experts, siloed
    - Inconsistent controls
- **Level 2 Repeatable:** Assessment criteria – Existence of role definition and policy documentation. Tool sets
    - Centralised tool usage
    - Roles are defined
    - Processes no longer dependent on experts
    - Master and Reference data beginning to be recognised
- **Level 3 Defined:** Assessment criteria – Existence of data management policies, scalable processes, consistent models and system controls
    - Emerging data management capability
    - Scalable DM Processes across organisation
    - DM recognised as an organisational enabler
    - Increase in DQ
    - Reduction in manual intervention
    - Process outcomes more predictable
- **Level 4 Managed:** Assessment criteria – Project success metrics, systems operational metrics, DQ metrics
    - Standardises tools for Data Management
    - Centralised planning and governance
    - Measurable increase in DQ and organisation wide capabilities such as data audits
- **Level 5 Optimisation:** Assessment criteria – Change management artefacts and process improvement metrics

- Highly predictable processes
- Focus on continuous improvement
- Tools enable visibility across processes
- Proliferation of data is controlled
- Well understood metrics used to measure quality and processes

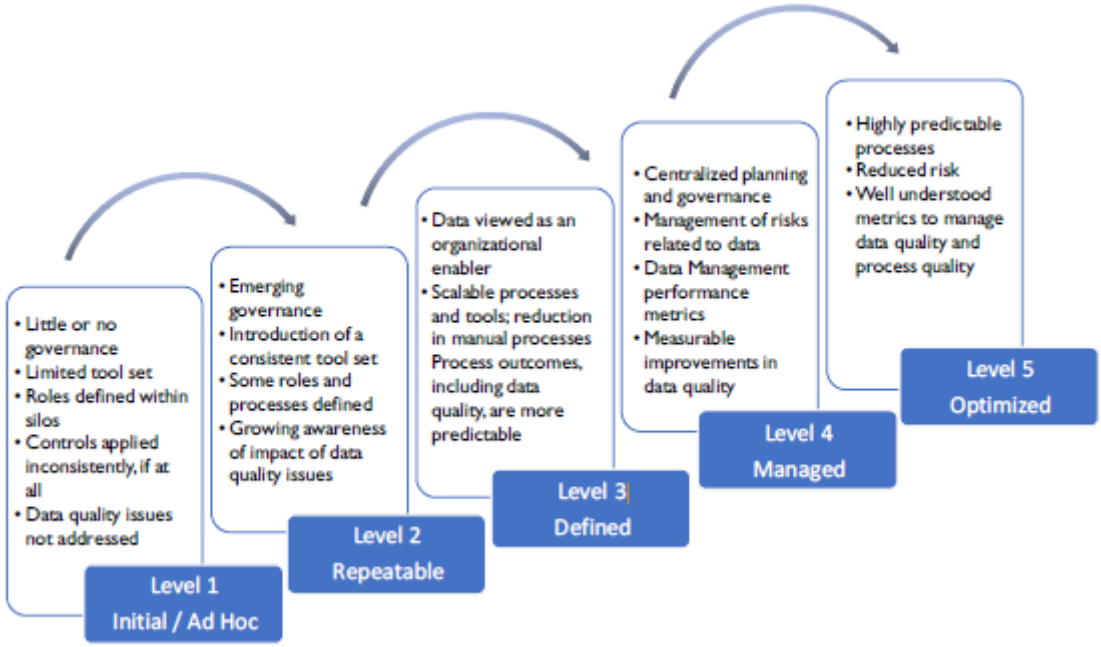

• Little or no governance
• Limited tool set
• Roles defined within silos
• Controls applied inconsistently, if at all
• Data quality issues not addressed
Level 1
Initial / Ad Hoc
• Emerging governance
• Introduction of a consistent tool set
• Some roles and processes defined
• Growing awareness of impact of data quality issues
Level 2
Repeatable
• Data viewed as an organizational enabler
• Scalable processes and tools; reduction in manual processes Process outcomes, including data quality, are more predictable
Level 3
Defined
• Centralized planning and governance
• Management of risks related to data
• Data Management performance metrics
• Measurable improvements in data quality
Level 4
Managed
• Highly predictable processes
• Reduced risk
• Well understood metrics to manage data quality and process quality
Level 5
Optimized

Figure 104 Data Maturity Model Example

1.3.2 Assessment Criteria

Assessment criteria are dependent on the process being evaluated. Will be evaluated along a scale such as 1 -Not started, 2 – In process, 3 – Functional, 4 – Effective, showing progress and movement to the next level.

When assessing using a model that can be mapped to a DAMA-DMBOK Data Management Knowledge Area, formulate criteria according to the relevant Context Diagram:

- **Activity:** To what degree is the activity in place? Are execution criteria defined? How well defined and executed is the activity? Are best practice outputs produced?
- **Tools:** To what degree is automation and supported by tools? Is tool training provided? Are tools available where needed? Optimally configured? To what extent is long term technology planning in place?
- **Standards:** To what degree is the activity supported by a common set of standards? Are standards enforced and supported by governance and management?
- **People and Resources**: To what degree is the organisation staffed to carry out the activity? What specific skills, training and knowledge are required? How well are roles and responsibilities defined?

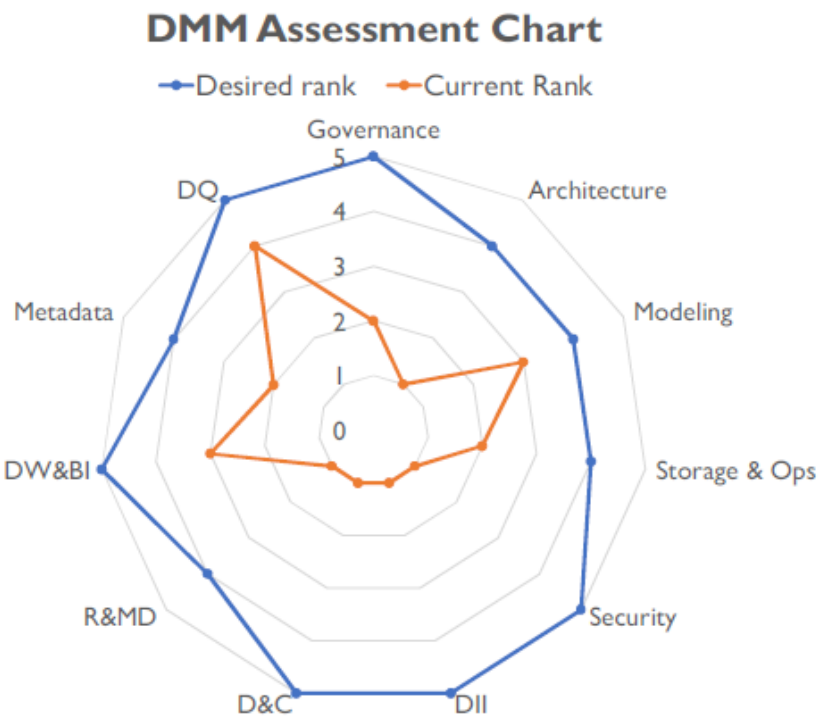


Figure 105 Example of a Data Management Maturity Assessment Visualization

1.3.3 Existing DMMA Frameworks

- **CMMI (Capability Maturity Model Institute) Data Management Maturity Model (DMM):** Provides assessment criteria for the following data management areas:
  - Data Management Strategy
  - Data Governance
  - Data Quality
  - Platform and Architecture
  - Data Operations
  - Supporting Procedures

  Identifies sub-processes within the areas and accounts for relation between the areas.

- **EDM Council DCAM:** Enterprise Data Management Council – financial services in USA – Data Management Capability Assessment Model
- **IBM Data Governance Council Maturity Model:** Organised around four key categories:
  - **Outcomes:** Data risk management and compliance, value creation
  - **Enablers:** Organisational structure awareness, policy, stewardship
  - **Core disciplines:** Data Quality Management, information lifecycle management, information security and privacy
  - **Supporting Disciplines:** Data Architecture, classification and Metadata, audit information, logging and reporting
- **Stanford Data Governance Maturity Model:** Focusses on Data Governance
- **Gartner's Enterprise Information Management Maturity Model**

2 Activities

2.1 Plan Assessment Activities

Plan for assessments to be conducted in a short, defined timeframe to expose current strengths and opportunities for improvement.

The goal is to reach a consensus view of current capabilities by interviewing business, data management and Information technology participants, supported by evidence.

2.1.1 Define Objectives

Must describe the focus and influence the scope of the assessment. Must be understood by business who help align with the organisation's strategic direction. Assessment objectives help decide which model to adopt and which business areas to prioritise.

2.1.2 Choose a Framework

In the context of assumptions about current state, and which will inform the organisation.

2.1.3 Define Organisational Scope

Keep scope manageable and phase into a larger enterprise assessment: Trade-offs between local and enterprise assessments:

- **Localised assessments:** Deeper into details, and is done more quickly. Can be aggregated and weighted to form an enterprise assessment
- **Enterprise Assessments:** Broad. parts of the organisation are sometimes disconnected. Can evaluate different functions based on the same criteria

2.1.4 Define Interaction Approach

Follow recommendations on the selected model. Information gathering should work well within the organisation's culture and minimise participants' time. Includes inspection and evaluation of artefacts and other evidence.

2.1.5 Plan Communications

Communications contributes to success. Ensure participants understand the assessment model, as well as how the findings will be used. Describe:

- The purpose of the DMMA
- How it will be conducted
- Their involvement
- Schedule of assessment activities

Include report on findings to all levels.

2.2 Perform Maturity Assessment

- **Gather Information:** Formal ratings of assessment criteria
- **Perform the Assessment:**
  - Review results against a rating method and assign a preliminary rating
  - Document supporting evidence
  - Review with participants to get consensus on a final rating
  - Document finding using model criteria statements
  - Visualisations to illustrate findings

2.3 Interpret Results

Identify improvement opportunities aligned with organisational strategy. Recommend actions, the steps towards the target state.

Present the meaning of the ratings to the organisation first. Link current capabilities with the business processes and strategies they support, and the benefits of improving these capabilities.

- **Report assessment results:**
  - Business drivers for assessment
  - Overall results of assessment
  - Ratings by topic with gaps indicated
  - Recommended approach to close gaps
  - Strengths of organisation observed
  - Risks to progress
  - Investment and outcomes options
  - Governance and metrics to measure progress
  - Resource analysis and potential future utilisation
  - Artefacts that can be reused
- **Develop executive briefings**
  - Prepare executive briefings that summarise findings – strengths, gaps and recommendations

2.4 Create a Targeted Program for Improvements

Recommendations from the DMMA should have a direct impact on data strategy and IT governance, and should be actionable.

Identify Actions and create a Roadmap:

- Sequenced activities for improvements in different DM functions
- Timeline
- Expected improvements in DMMA ratings
- Oversight activities

Should be accompanied by an approach for measuring progress.

2.5 Re-assess Maturity

Reassessments are part of a lifecycle of continuous improvement.

- First assessment establishes a baseline
- Define re-assessment parameters (including organisational scope)
- Repeat DMMA on a published schedule
- Track trends relative to initial baseline
- Develop recommendations based on findings

3 Tools

- **Data Management Maturity Framework:** Primary tool
- **Communication plan:** Engagement model for stakeholders, type of information required and schedule
- **Collaboration tools:** Allow findings to be shared

- **Knowledge Management and Metadata Repositories:** Data standards, policies, methods, agendas, minutes of meetings or decisions and business and technical artefacts are managed here

4 Techniques

4.1 Selecting a DMM Framework

- **Accessibility:** Non-technical terms that convey functional essence of the activity
- **Comprehensiveness:** Framework address a broad scope of data management activities
- **Extensible and flexible:** Able to enhance model. Model can be used in part or whole
- **Future progress path built-in:** Logical way forward within each of the DMM functions
- **Industry-agnostic vs. industry-specific:** Depends on the organisation, but best practices should be adhered to
- **Level of abstraction or detail:** To be related to the organisation and the work it performs
- **Non-prescriptive:** Describes what must be performed, not how
- **Organised by topic:** DM activities placed in context
- **Repeatable:** Consistent over time
- **Supported by a neutral, independent organisation:** Vendor neutral to avoid conflicts of interest
- **Technology neutral:** Based on practices rather than tools
- **Training support included:** Model is supported by comprehensive training

4.2 DAMA-DMBOK Framework Use

DAMA-DMBOK can be used to establish criteria for a DMMA.

5 Guidelines for a DMMA

5.1 Readiness Assessment / Risk Assessment

Risk	Mitigation
Lack of organizational buy-in	Socialize the concepts related to the assessment. Establish benefit statements before conducting the assessment. Share articles and success stories. Engage an executive sponsor to champion the effort and review the results.
Lack of DMMA expertise Lack of time or in-house expertise Lack of communication planning or standards	Use third party resources or specialists. Require knowledge transfer and training as part of the engagement.
Lack of 'Data Speak' in the organization; Conversations on data quickly devolve into discussions about systems	Relate the DMMA to specific business problems or scenarios. Address in the communications plan. The DMMA will educate all participants regardless of background and technical experience. Orient participants to key concepts prior to the DMMA.
Incomplete or out-of-date assets for analysis	Flag 'as of' or balance the rating accordingly. For example, give a -1 to everything that is over 1 year out-of-date.
Narrow focus	Reduce the investigation depth to a simple DMMA and go to other areas for a quick assessment to establish ratings for a later comparative baseline. Conduct the first DMMA as a pilot, then apply lessons learned to address a broader scope. Present in-scope focus of proposed assessment in context of DAMA-DMBOK Knowledge Areas. Illustrate what is being left out of scope and discuss the need to include.
Inaccessible staff or systems	Reduce the horizontal scope of the DMMA by focusing only on available Knowledge Areas and staff
Surprises arise such as a regulation changes	Add flexibility into the assessment work stream and focus.

Table 33 Typical Risks and Mitigations for a DMMA

5.2 Organisational and Cultural change

DMMA locates the organisation on a maturity scale and provides a roadmap for improvement, pointing the organisation through change. DMMA can provide a shared vision by coalescing different perspectives. Needs to be supported by effective governance.

6 Maturity Management Governance

DMMA is part of an overall set of data governance activities, each having a lifecycle. DMMA Lifecycle:

- Initial planning and assessment
- Recommendations
- Action plan
- Periodic re-evaluation

6.1 DMMA Process Oversight

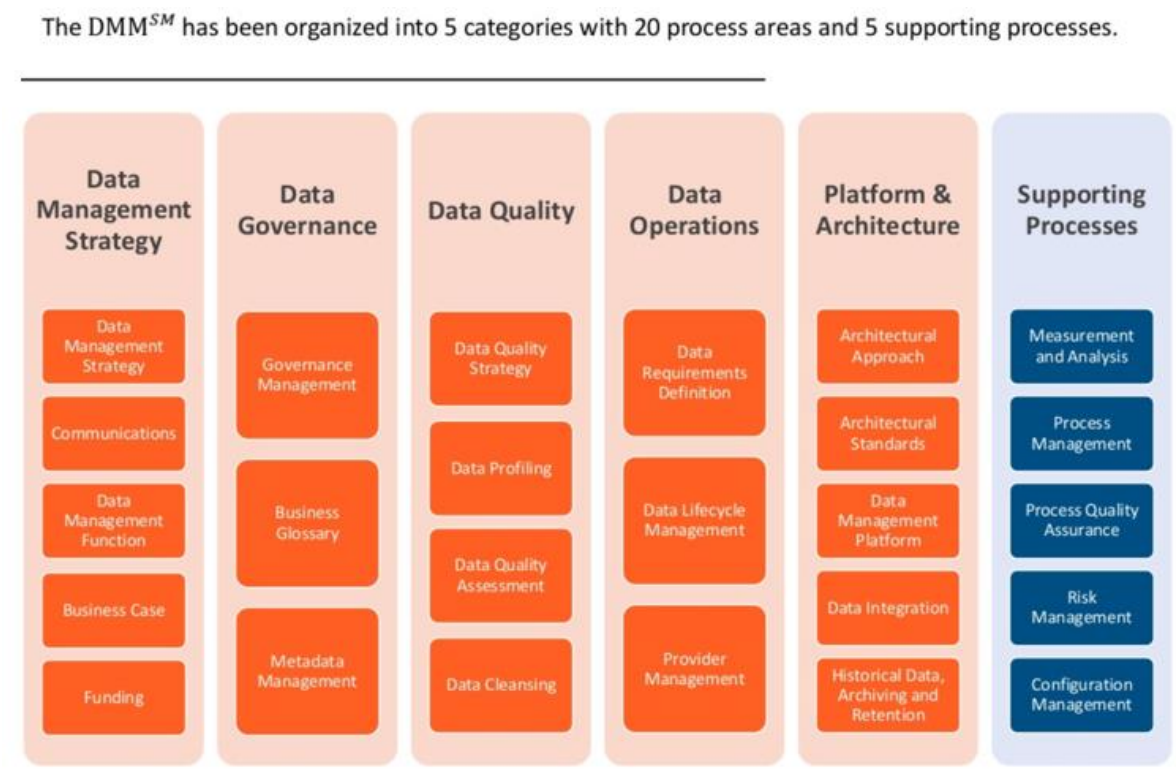
Data Governance team

6.2 Metrics

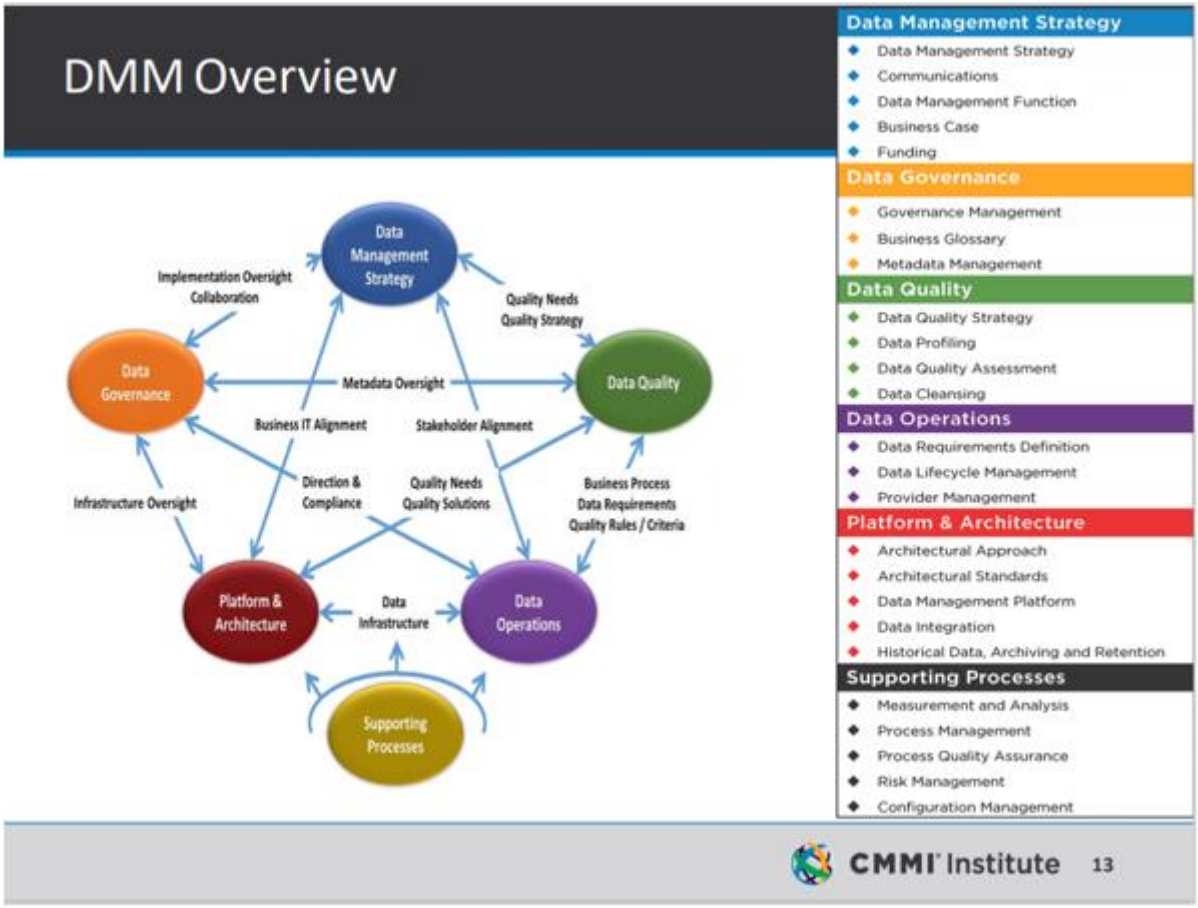
Represent the current state:

- **DMMA Ratings**: Snapshot of the organisations capability level
- **Resource utilisation rates:** Cost management in the form of a head count

- **Risk exposure:** Or ability to respond to risk
- **Spend management:** How the cost of data management is spread over the organisation. Overlap with Data Governance metrics
    - Data management sustainability
    - Achievement of initiative goals and objectives
    - Effectiveness of communication
    - Effectiveness of education and training
    - Speed of change adoption
    - Data management value
    - Contributions to business objectives
    - Reductions in risks
    - Improved efficiency in operations
- **Inputs to the DMMA**: Indicate completeness of coverage
- **Rate of change**: Baseline is established. Periodic reassessment to trend improvement



CMMI-DMM from the CMMI Institute



Recognizing Business Data Capability Levels

Disabled	Enabled	Led	Driven
• DISTRUST • People Story Telling • Undefined Sources for Decision Making	• AWARENESS • Augmented Story Telling • Curated Data Sources for Decision Making	• INSIGHTS • Data Story Telling • Data Estate Sources for Decision Making	• ACTION • Feedback Story Telling • Data Value Chain Source of Decision Making

From the Course How-to Govern Data by Howard Diesel

Chapter 16
Data Management Organisation and Role Expectations

1 Introduction

The set of principles to consider when establishing a Data Management or Data Governance Organisation.

Awareness, Ownership and Accountability are the keys to activating and engaging people in data management initiatives, policies and processes.

2 Understanding Existing Organisation and Cultural Norms

Before defining any new organisation or improvements to the existing one, it is important to understand the current state, relating to culture, the existing operating model and people.

- **The role of data in the organisation:** Data driven processes, data requirements understood, how well recognised is the role data plays in the organisation?
- **Cultural norms about data:** Potential cultural obstacles
- **Data management and data governance practices:** Who does the work? How? Who makes decisions?
- **How work is organised and executed:** Project-focussed and operational execution. Committees to support the data management initiative
- **How reporting relationships are organised:** Centralised, decentralised, hierarchical or flat
- **Skill levels:** Levels of data knowledge and data management knowledge of SMEs and from line staff to executives

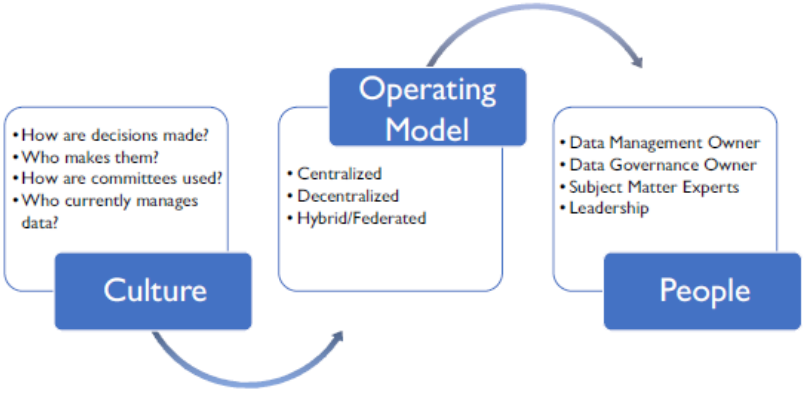


Figure 106 Assess Current State to Create an Operating Model

Assess the level of satisfaction with the current state:

- Does the organisation have the information it needs to make sound, timely business decisions?
- Does the organisation have confidence in its revenue reports?
- Can it track the organisational KPIs?

- Compliance with data management laws?

Most organisations crafting a Data Management Organisation are in the middle of the maturity scale.

A Data Management Organisation should align with the company's organisational hierarchy and resources. Cross-functional participation from various stakeholders:

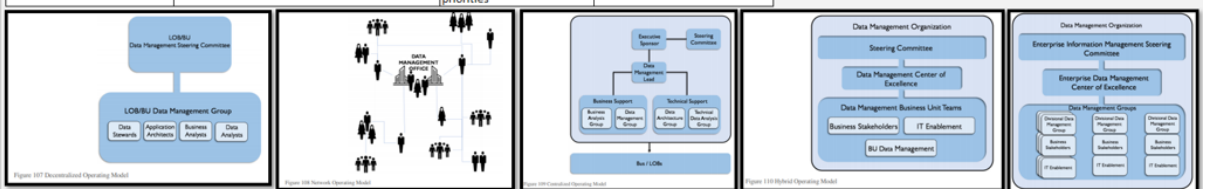
- Involve employees currently performing data management functions first
- Determine how much change is required to improve the existing data management methods
- Roadmap the changes from an organisational perspective

3 Data Management Organisational Constructs

The operating model is a framework articulating roles, responsibilities and decision making process. How people and functions will collaborate. Facilitates communication and a process to resolve issues.

Organization Construct	Central Tenant	Pro	Con
Decentralized Operating Model	DM Responsibilities are distributed across different lines of business & IT. No single owner - collaboration	Relatively flat structure Alignment of data management to lines of business or IT	Many Participants involved with governance bodies and in decision-making
Network Operating Model	Decentralized informality made more formal through a documented series of connections and RACI definitions	Flat Structure Busniess Alignment Quick set up Accountability (no impact on Org Structure)	Maintain & Enfroce Expecations related to RACI
Centralized Operating Model	Everything is owned by the Data Management Organization. Responsible for all DM Knowledge Areas & Data Ownership	Formal Executive Position for Data Management CDO Organization Simpler Decision-Making & Accountability	Formal separation of the role of Data Management from Business Processes and Departments Requires a new Organization
Hybrid Operating Model	Benefits of Decentralized & Centralized Models. Centralized CoE works with Decetranlized Business Unit Groups	Establish appropriate direction from the top. Accontable Executive Business Unit can align to business priorities	Requires Set Up Managing different priorities with one CoE
Federated Operating Model	Centralized Strategy with decentralized execution	Prioritize based on specific data entities, divisional challenges, or regional priorities	Complexity - lots of layers

Organizational Constructs Summary



3.1 Decentralised Operating Model

Data management responsibilities distributed over lines of business and IT. Collaboration is committee based and there is no single owner.

Pros:

- Flat structure
- alignment of data management to lines of business or IT
    - understand data requirements clearly
- Relatively easy to implement or improve

Cons:

- Many participants in decision making
- Harder to implement collaborative decisions than centralised edicts
- Less formal therefore harder to sustain over time
- Difficult to coordinate

- Difficult to define ownership

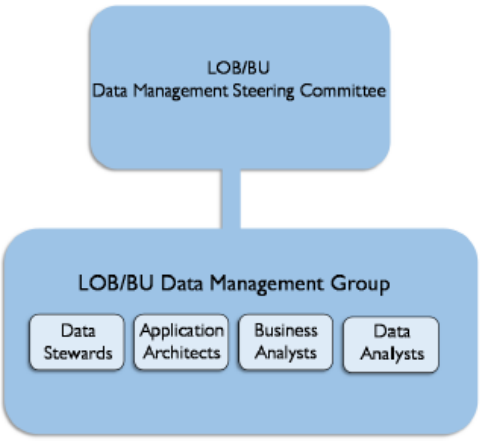


Figure 107 Decentralised Operating Model

3.2 Network Operating Model

Formalise through a RACI (Responsible, Accountable, Consulted and Informed) matrix. Operates as a series of known connections between people and roles.

Pros:

- Flat structure
- Alignment
- quick set-up
- RACI creates accountability

Cons:

- Need to maintain and enforce expectations related to the RACI

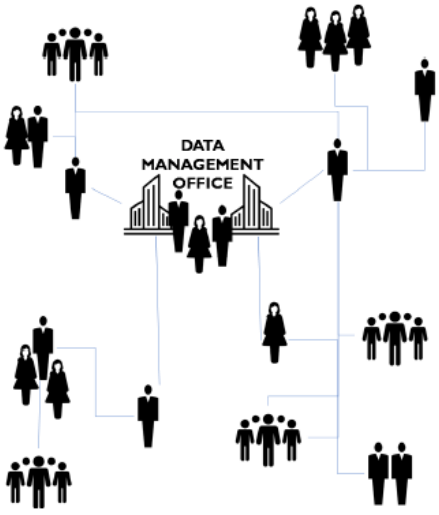


Figure 108 Network Operating Model

3.3 Centralised Operating Model

The most formal and mature data management operating model. Everything is owned by the Data Management Organisation, and everyone reports to the responsible data management leader.

Pros:

- A formal executive position for data management or data governance
- Easy decision making because accountability is clear
- Data can be managed by type or subject area

Cons:

- Significant organisational change to implement
- Formalisation of the data management roles can result in separation from business processes and knowledge loss over time

The leader should report to business (CDO), not IT (CIO).

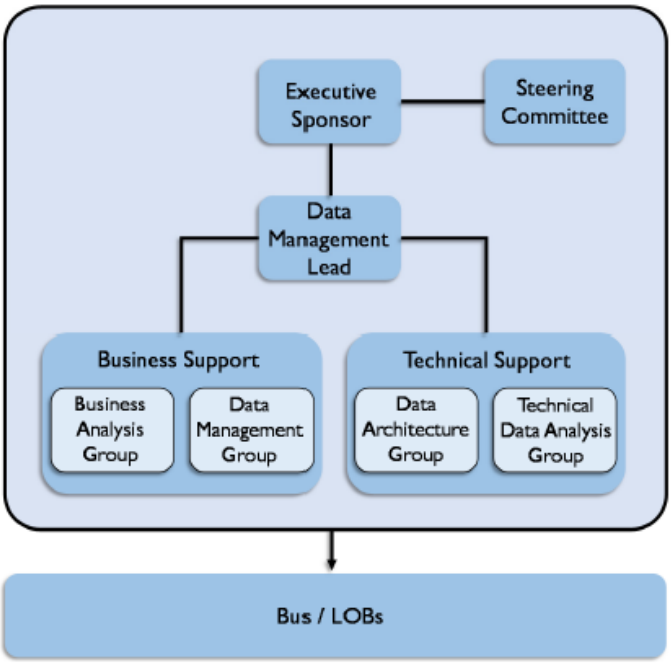


Figure 109 Centralised Operating Model

3.4 Hybrid Operating Model

Benefits of the centralised and decentralised models. Centralised data management Centre of Excellence works with decentralised business unit groups. Some groups remain decentralised.

Pros:

- There is an executive accountable for data management and/or data governance
- Business units have broad accountability
- There is support from the dedicated Centre of Excellence to help with changes

Cons:

- Challenge to set up, Centre of Excellence needs dedicated staff
- Business unit teams have different priorities
- Conflicts between priorities of central organisation and the decentralised organisation

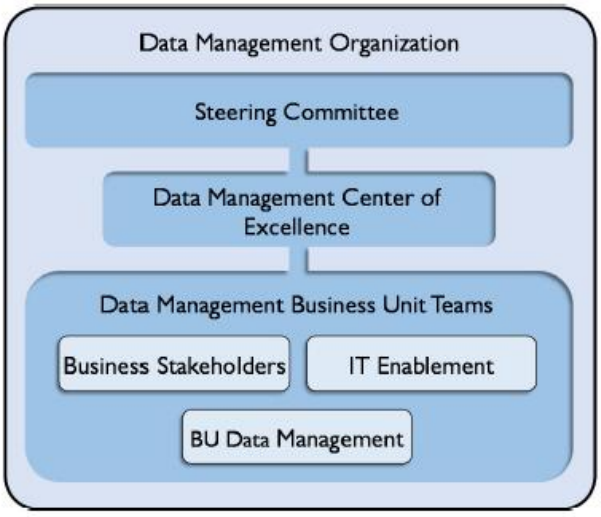


Figure 110 Hybrid Operating Model

3.5 Federated Operating Model

Centralised strategy with decentralised execution, good for large organisations.

Pros:

- Federation helps the organisation prioritise based on specific data entities, divisional challenges or regional priorities

Cons:

- Complex – lots of layers
- Needs balance between autonomy in lines of business and needs of the enterprise

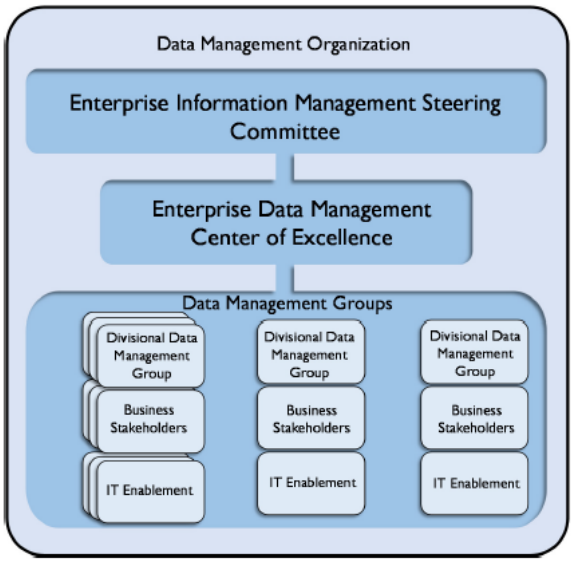


Figure 111 Federated Operating Model

3.6 Identifying the Best Model for an Organisation

Assess organisational structure, how decisions are made and implemented. This should place the organisation somewhere on the spectrum between decentralised and centralised.

3.7 DMO Alternatives and Design Considerations

Tips to keep in mind when constructing an operating model:

- Determine starting point by assessing current state
- Tie the operating model to the organisational structure
- Take into account:
  - Organisation complexity + maturity
  - Domain complexity + maturity
  - Scalability
- Must get executive sponsorship
- Ensure any leadership forum is decision making
- Consider pilot programs and waves of implementation
- Focus on high value, high impact data domains
- Use what already exists
- Never take one-size-fits-all approach

4 Critical Success Factors

1. **Executive sponsorship:** Must be able to engage other leaders in support of changes
2. **Clear Vision:** Plan to drive the Data Management program with understanding
3. **Proactive Change Management:** Addresses people challenges
4. **Leadership alignment:** Agreement and unified support for the Data Management program
5. **Communication:**
   - Start early and continue openly and often
   - Create a story around the Data Management initiative
   - Consistent messages – importance of Data Management
   - Customised to stakeholder groups
   - Repeated and tested over time
6. **Stakeholder engagement:** Stakeholder analysis determines best way to handle different stakeholders
7. **Orientation and training:** Focussed training allows people to perform their roles effectively
8. **Adoption measurement:** Measure:
   - Adoption
   - Amount of improvement
   - How well does data management influence solutions with measurable results?
   - Improves processes, projects
   - Improves identification and reaction to risk
   - Innovation- how well does data management fundamentally change how business is conducted?
   - Trusted analysis
9. **Adherence to Guiding Principles:** These are shared organisational values which serve as the guiding points from which decisions are made
10. **Evolution not Revolution:** Helps minimise big changes in high risk areas. Better to evolve and mature over time

5 Build the Data Management Organisation

5.1 Identify Current Data Management Participants

Start with teams already engaged in DM activities. Review organisation for other data people fulfilling the roles and responsibilities. Identify gaps. Ensure roles are assigned to the right people at the right levels to ensure decision making credibility.

5.2 Identify Committee Participants

It is best to start a well informed and focussed Data Governance Steering Committee.

- Data Governance Steering Committee and working groups
- Get the right people on the committee
- Keep them well informed and focussed
- Start a new committee so that DM gets the attention it requires
- Conduct stakeholder analysis to identify executive sponsors

5.3 Identify and Analyse Stakeholders

A stakeholder is any person or group who can influence or be affected by the Data Management Program. Can be internal or external.

Stakeholder analysis answers questions:

- Who will be affected by data management?
- How will roles and responsibilities shift?
- How might those who are affected respond to changes?

What issues and concerns will people have?

Critical stakeholders need to be brought along. Consider:

- Who controls critical resources?
- Who can block the data management initiatives?
- Who can influence other critical constituents?
- How supportive stakeholders are of the upcoming changes?

5.4 Involve the Stakeholders

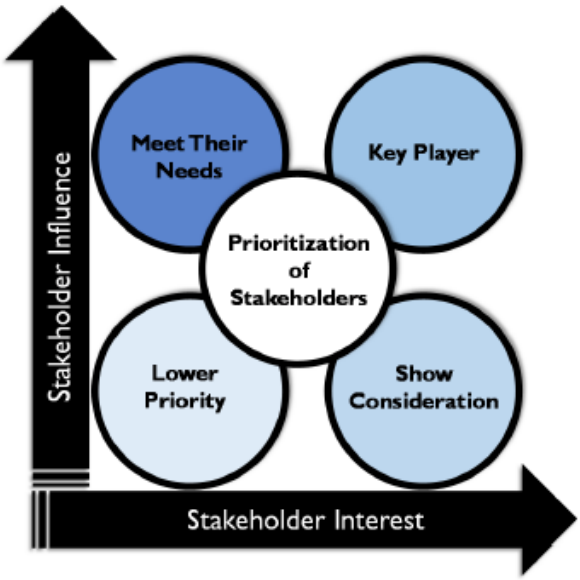


Figure 112 Stakeholder Interest Map

Person driving DM effort must articulate why each stakeholder should be involved.

To ensure long-term support, Stakeholders and Executive Sponsor need to understand:

- Why they are necessary to the success of the DM program
- The link between DM output and their personal and professional goals

6 Interactions Between the DMO and other Data-oriented Bodies

- **The Chief Data Officer:** Required to bridge the gap between business and IT. Mandate is:
  - Establish an organisational data strategy
  - Align data-centric requirements with IT and business resources
  - Establish data governance standards, policies and procedures
  - Provide advice to business for data-dependent initiatives
  - Evangelise the importance of good information management principles to business stakeholders internally and externally
  - Oversight of data usage in analytics and BI
  - The DMO reports upwards through CDO
- **Data Governance:** Data Governance provides the framework for the DM Program to align with business priorities
  - John Ladley: Data Governance is about “doing the right things” and Data Management is about “doing things right”
  - Centralised model: Data Governance Office can report to the Data Management Organisation or vice versa
  - Decentralised Model: a tight partnership between the DGO and DMO
- **Data Quality:** Data Quality Management is a key component of a Data Management Organisation.
  - Many DMOs come out of a desire to improve Data Quality. Can start in business lines but a mature data quality practice benefits from a unified Centre of Excellence

    - DQ program is usually decentralised, network or hybrid
    - Expands to Master, Reference and Metadata Management
    - Align DQ operating model to overall DMO for standardisation and resource sharing
- **Enterprise Architecture:** Designs the master blueprints for an organisation to meet its strategic objectives. Includes:
    - Technology Architecture
    - Application Architecture
    - Information (Data) Architecture – a key capability of an effective Data Management Organisation
    - Business Architecture

  Data Architects can sit in Enterprise architecture or in the DMO and interface with other architects via Architecture Review Boards (ARB) which review projects and give guidance on adherence to architectural standards.

  When there are no Data Architects, Data Management interfaces with architecture in a few ways:

    - Through Data Governance
    - Through the ARB
    - Ad-hoc
- **Managing a Global Organisation:** Time zone challenges, country specific laws and physical distances, therefore need to pay specific attention to:
    - Adhering to standards
    - Synchronising processes
    - Aligning accountability
    - Training and communication
    - Monitoring and measuring effectively
    - Developing economies of scale
    - Reducing duplication effort

  Networked or Federated models work best for global organisations

7 Data management Roles

All IT roles can be mapped to points on the data lifecycle. Business roles create, access or manipulate data.

7.1 Organisational roles

IT Data Management Organisations provide data, application, technical architecture and DBA services.

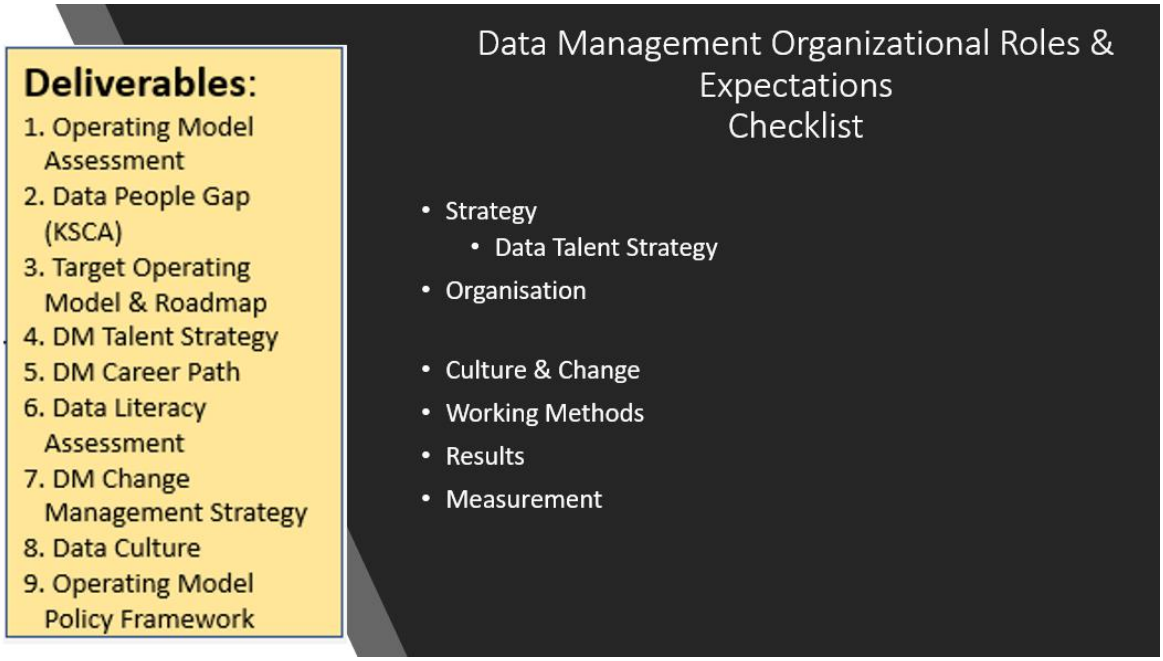
- **Centralised Data Management Services Organisation:** Focussed solely on Data Management. Team: DM Executive and other DM Managers, Data Architects, Data Analysts, Data Quality Analysts, DBAs, Data Security administrators, Metadata Specialists, Data Modellers, Data Administrators, Data Warehouse Architects, Data Integration Architects and BI Analysts
- **Federated Data Management Service:** IT units each focussed on a facet of Data Management

- **Hybrid Data Management Service:** Each business function has own developers, but DBA is centralised

Business functions focussed on data management are mostly associated with Data Governance of Enterprise Information Management teams.

7.2 Individual Roles

- **Executive roles:** CIO and CTO in IT or CDO in business
- **Business Roles:** Data Stewards, Business Process Analysts and Process Architects, business consumers of data
- **IT Roles:**
    - **Data Architect:** Senior analyst responsible for data architecture and data integration. Enterprise or function level
    - **Data Modeller:** Responsible for capturing and modelling data requirements, data definitions, business rules, data quality requirements, and logical and physical data models
    - **Data Model Administrator:** Responsible for model version control and change control
    - **Database Administrator:** Responsible for the design, implementation and support of structured data assets and the performance of the technology that makes data accessible
    - **Data Security Administrator:** Responsible for ensuring controlled access
    - **Data Integration Architect:** Senior data integration developer responsible for designing technology to integrate and improve quality of enterprise data assets
    - **Analytics/Report Developer:** Responsible for creating reporting and analytical application solutions
    - **Application Architect:** Senior developer responsible for integrating application systems
    - **Technical Architect:** Senior technical engineer responsible for coordinating and integrating the IT infrastructure and the IT technology portfolio
    - **Technical Engineer:** Senior technical analyst responsible for researching, implementing, administering and supporting a portion of the IT infrastructure
    - **Help Desk Administrator:** Handling, tracking and resolving IT issues
    - **IT Auditor:** Internal or external auditor of IT responsibilities, including data quality and data security
- **Hybrid Roles:**
    - **Data Quality Analyst:** Determine fitness of use of data and monitoring ongoing condition. Root cause analysis of data issues
    - **Metadata Specialist:** Integration, control and delivery of Metadata
    - **Business Intelligence Architect:** Senior BI analyst responsible for design of the BI user environment
    - **Business Intelligence Analyst/Administrator:** Supports effective use of BI data
    - **Business Intelligence Program Manager:** Integrates BI requirements and initiatives into a cohesive program and roadmap



Deliverables and Categories Checklist

DM Organization & Role Expectation : What?

Definition

Awareness, Ownership and Accountability are the keys to activating and engaging people in data management initiatives, policies and processes.

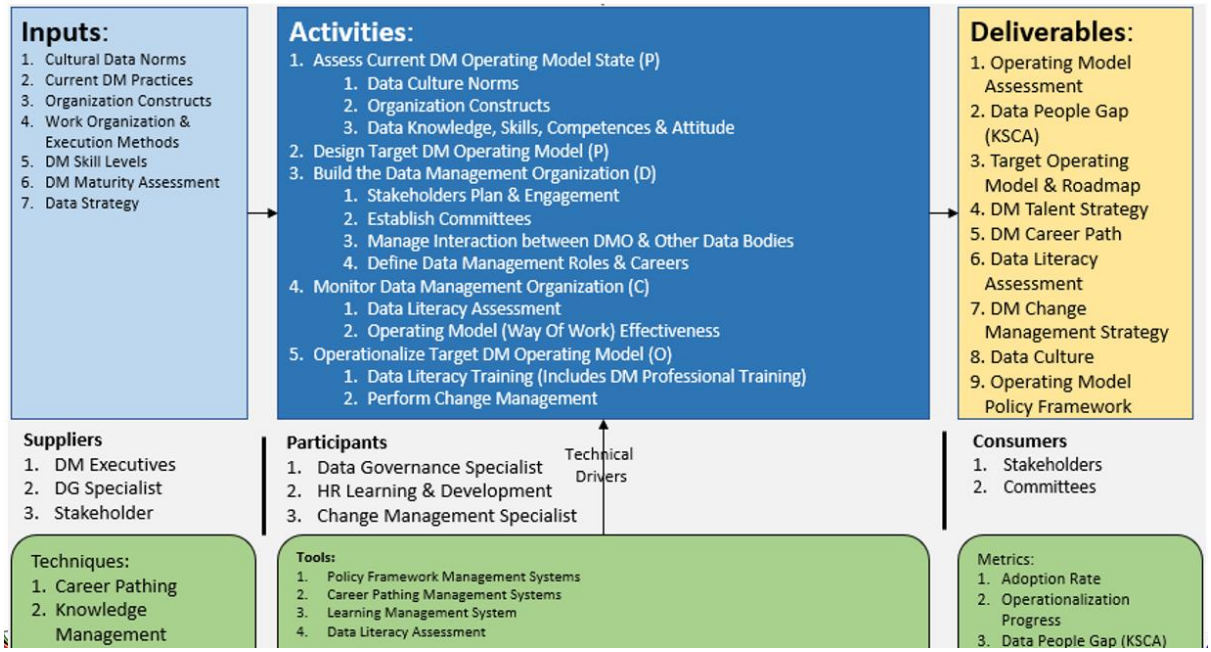
- Cultural Norms
- Organizational Constructs
- Data Knowledge, Skills, Competences & Attitudes

Definition of the Data Management Organisation and Role Expectation Chapter from the How-to Govern Data Course by Howard Diesel

DM Organization & Role Expectation: Why?

Goals	Business Drivers
• Ensure that the Role of Data In the Organization compliments the Business & Data Strategy • Ensure that DM Organizational Constructs are aligned & support the Business • Ensure that DM Participants have the appropriate KSCA levels	• Activate & Engage People In Data Management Initiatives, Policies & Processes • Efficient & Effective Data Management Processes to support Business Decision-Making

Goals and Business Drivers for the Data Management Organisation and Role Expectations Chapter from the How-to Govern Data Course by Howard Diesel


Inputs:
1. Cultural Data Norms
2. Current DM Practices
3. Organization Constructs
4. Work Organization & Execution Methods
5. DM Skill Levels
6. DM Maturity Assessment
7. Data Strategy
Activities:
1. Assess Current DM Operating Model State (P)
1. Data Culture Norms
2. Organization Constructs
3. Data Knowledge, Skills, Competences & Attitude
2. Design Target DM Operating Model (P)
3. Build the Data Management Organization (D)
1. Stakeholders Plan & Engagement
2. Establish Committees
3. Manage Interaction between DMO & Other Data Bodies
4. Define Data Management Roles & Careers
4. Monitor Data Management Organization (C)
1. Data Literacy Assessment
2. Operating Model (Way Of Work) Effectiveness
5. Operationalize Target DM Operating Model (O)
1. Data Literacy Training (Includes DM Professional Training)
2. Perform Change Management
Deliverables:
1. Operating Model Assessment
2. Data People Gap (KSCA)
3. Target Operating Model & Roadmap
4. DM Talent Strategy
5. DM Career Path
6. Data Literacy Assessment
7. DM Change Management Strategy
8. Data Culture
9. Operating Model Policy Framework
Suppliers
1. DM Executives
2. DG Specialist
3. Stakeholder
Participants
1. Data Governance Specialist
2. HR Learning & Development
3. Change Management Specialist
Technical Drivers
Consumers
1. Stakeholders
2. Committees
Techniques:
1. Career Pathing
2. Knowledge Management
Tools:
1. Policy Framework Management Systems
2. Career Pathing Management Systems
3. Learning Management System
4. Data Literacy Assessment
Metrics:
1. Adoption Rate
2. Operationalization Progress
3. Data People Gap (KSCA)

Context Diagram: The DM Organisation and Role Expectations (Howard Diesel)

Chapter 17
Data Management and Organisational Change Management

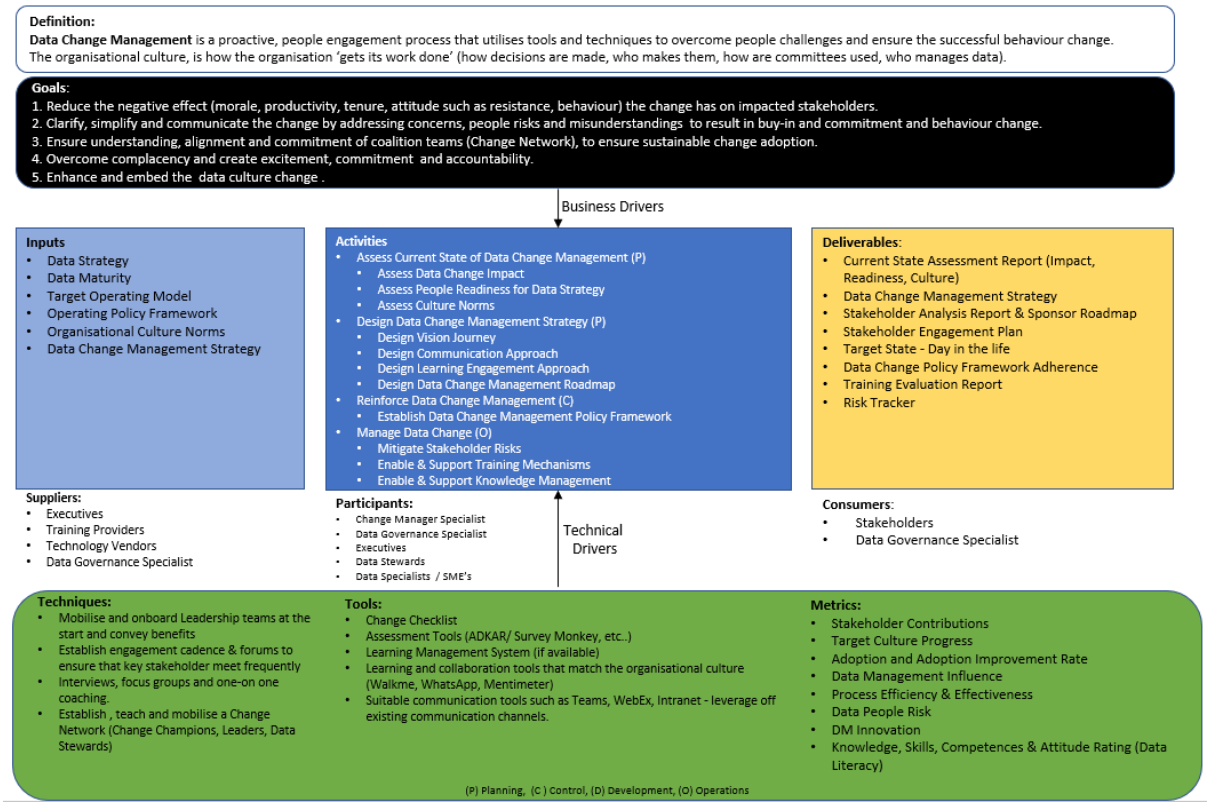
1 Introduction

Improving data management practices requires changing how people work together, how they understand the role of data on their organisations and how they use data and deploy technology to support organisational processes. Successful data management practices require:

- Manage on the horizontal by aligning accountabilities along the Information Value Chain
- Change focus from vertical accountability (silo) to shared stewardship
- Evolve information quality into a core value of the organisation
- Shift thinking about information quality from data cleansing to a fundamental organisational capability
- Processes to measure cost of poor data management and value of disciplined data management

Change is required at all levels. Critical to manage it to avoid dead-end initiatives, loss of trust and credibility damage. Understand:

- Why change fails
- The triggers for effective change
- The barriers to change
- How people experience change


Definition:
Data Change Management is a proactive, people engagement process that utilises tools and techniques to overcome people challenges and ensure the successful behaviour change.
The organisational culture, is how the organisation 'gets its work done' (how decisions are made, who makes them, how are committees used, who manages data).
Goals:
1. Reduce the negative effect (morale, productivity, tenure, attitude such as resistance, behaviour) the change has on impacted stakeholders.
2. Clarify, simplify and communicate the change by addressing concerns, people risks and misunderstandings to result in buy-in and commitment and behaviour change.
3. Ensure understanding, alignment and commitment of coalition teams (Change Network), to ensure sustainable change adoption.
4. Overcome complacency and create excitement, commitment and accountability.
5. Enhance and embed the data culture change .
Business Drivers
Inputs
Data Strategy
Data Maturity
Target Operating Model
Operating Policy Framework
Organisational Culture Norms
Data Change Management Strategy
Activities
Assess Current State of Data Change Management (P)
Assess Data Change Impact
Assess People Readiness for Data Strategy
Assess Culture Norms
Design Data Change Management Strategy (P)
Design Vision Journey
Design Communication Approach
Design Learning Engagement Approach
Design Data Change Management Roadmap
Reinforce Data Change Management (C)
Establish Data Change Management Policy Framework
Manage Data Change (O)
Mitigate Stakeholder Risks
Enable & Support Training Mechanisms
Enable & Support Knowledge Management
Deliverables:
Current State Assessment Report (Impact, Readiness, Culture)
Data Change Management Strategy
Stakeholder Analysis Report & Sponsor Roadmap
Stakeholder Engagement Plan
Target State - Day in the life
Data Change Policy Framework Adherence
Training Evaluation Report
Risk Tracker
Suppliers:
Executives
Training Providers
Technology Vendors
Data Governance Specialist
Participants:
Change Manager Specialist
Data Governance Specialist
Executives
Data Stewards
Data Specialists / SME's
Technical Drivers
Consumers:
Stakeholders
Data Governance Specialist
Techniques:
Mobilise and onboard Leadership teams at the start and convey benefits
Establish engagement cadence & forums to ensure that key stakeholder meet frequently
Interviews, focus groups and one-on one coaching.
Establish , teach and mobilise a Change Network (Change Champions, Leaders, Data Stewards)
Tools:
Change Checklist
Assessment Tools (ADKAR/ Survey Monkey, etc..)
Learning Management System (if available)
Learning and collaboration tools that match the organisational culture (Walkme, WhatsApp, Mentimeter)
Suitable communication tools such as Teams, WebEx, Intranet - leverage off existing communication channels.
Metrics:
Stakeholder Contributions
Target Culture Progress
Adoption and Adoption Improvement Rate
Data Management Influence
Process Efficiency & Effectiveness
Data People Risk
DM Innovation
Knowledge, Skills, Competences & Attitude Rating (Data Literacy)
(P) Planning, (C) Control, (D) Development, (O) Operations

Context Diagram: Data Management Organizational Change Management

DM Organisational Change Management: Why?

Goals	Business Drivers
• Reduce the negative effect (morale, productivity, tenure, attitude such as resistance, behaviour) the change has on impacted stakeholders • Clarify, simplify and communicate the change by addressing concerns, people risks and misunderstandings to result in buy-in and commitment and behaviour change • Ensure understanding, alignment and commitment of coalition teams (Change Network), to ensure sustainable change adoption • Overcome complacency and create excitement, commitment and accountability • Enhance and embed the data culture change	• Supporting the Data Vision by ensuring that people and culture are "on the bus" • Get the "Right People" on the bus and in the "Right Seats" • Ensure that people feel supported by the organization and data • Get the wrong people off the bus

Goals and Business Drivers (Howard Diesel)

2 Laws of Change

Recognise early why change is not easy:

- **Organisations don't change, people change:** People behave differently because they recognise the value in doing so

- **People don't resist change. They resist being changed**: More likely to change if they have been engaged in defining the change
- **Things are the way they are because they got that way**: Understand the origins of current data management practices and avoid past mistakes
- **Unless there is push to change, things will likely stay the same**: Something must be done differently
- **Change would be easy if it wasn't for all the people**

Change requires Change Agents who pay attention to the people and catch problems before they arise.

Ultimately change requires a clear VISION of change goals which are communicated vividly and frequently to stakeholders to get buy-in and continued support when challenges arise.

3 Not Managing a Change: Managing a Transition

William Bridges: Change is a transition moving through three phases: The Ending, The Neutral Zone and The New Beginning. Many changes fail because the Ending and Neutral Zone are not recognised and managed.

All individuals go through the phases at their own pace. Bridges emphasises the first task of the change manager is to understand the destination (vision). The ultimate goal of Change Management is to convince people they need to start the journey.

Transition Phase	Description
The Ending	• When we acknowledge that there are things we need to let go of. • When we recognize that we have lost something. • Example: Changing jobs – even when an individual chooses to change jobs, there are still losses such as losing close working friends.
The Neutral Zone	• When the old way has finished but the new way isn't here yet. • When everything is in flux and it feels like no one knows what they should be doing. • When things are confusing and disorderly. • Example: Moving to a new house. The first few days or even months after moving, the new house is not home yet and things are quite probably in turmoil.
The New Beginning	• When the new way feels comfortable, right, and the only way. • Example: Having a baby. After a few months in the neutral zone of turmoil, you come to a stage when you cannot imagine life without your new baby.

Table 35 Bridges's Transition Phases

The role of the Change Agent (manager or leader) is to help people realise that the stages of a transition are perfectly natural.

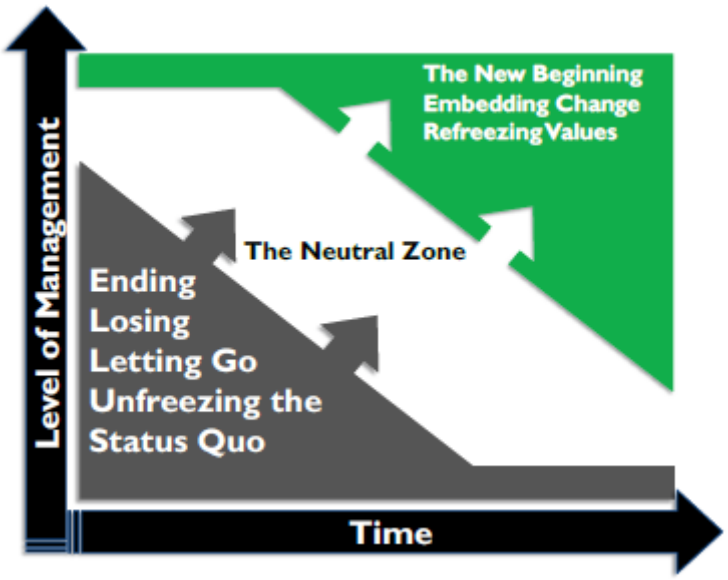


Figure 113 Bridges's Transition Phases

Key points managers should be aware of:

- The Ending:
  - Help everyone understand current problems and why change is necessary
  - Identify who will lose what
  - Accept that losses are subjective, and perceived differently. Don't argue
  - Acknowledge grieving sympathetically
  - Define what is over and what is not
  - Treat the past with respect
  - Show how ending something ensures things that matter to people are continued and improved
  - Give people information
  - Use stakeholder analysis to map best approaches for different individuals
- The Neutral Zone:
  - Recognise it is a difficult phase, but that everyone must go through it
  - Get people working together and experimenting
  - Help people feel they are still valued
  - Praise people with good ideas. PDSA (Plan, Do, Study, Act) model encourages trying things out and learning from the process
  - Keep giving people information
  - Provide feedback on results of ideas being tested and decisions made
- The New Beginning
  - Do not force a beginning before its time
  - Ensure people know their part in the new system
  - Make sure policies and procedures are clear
  - Plan to celebrate the new beginning
  - Keep giving people information

4 Kotter's Eight Errors of Change Management

John P Kotter – *Leading Change* – identified eight reasons why organisations fail to execute change.

4.1 Error #1: Allowing Too Much Complacency

Do not plunge ahead without establishing a sense of urgency. Related to the need to drive up dissatisfaction with the status quo identified in the Gleicher formula (See section 6)

Example Scenario	How it might manifest
Response to a Regulatory Change	"We're OK. We haven't been fined under the current rules."
Response to Business Change	"We've been supporting the business successfully for years. We'll be OK."
Response to Technology Change	"That new technology is unproven. Our current systems are stable and we know how to work around issues."
Response to Problems or Errors	"We can assign a troubleshooting team to that and patch the issues up. There are bound to be some people available in [Insert name of Department or Team here]."

Table 36 Complacency Scenarios

4.2 Error #2: Failing to Create a Sufficiently Powerful Guiding Coalition

There needs to be commitment from the very top leaders. A Guiding Coalition is a powerful and enthusiastic team of volunteers from across the organisation that helps put new strategies into effect.

4.3 Error #3: Underestimating the Power of Vision

A vision is a clear and compelling statement of where the change is leading. Vision needs to be communicated by connecting with people.

Often the change is communicated as a new system or technology. The change and vision are actually what the organisation can achieve with the new technology.

4.4 Error #4: Under Communicating the Vision by a Factor of 10, 100 or 1000

Kotter recommends using words and deeds to communicate the vision.

4.5 Error #5: Permitting Obstacles to Block the Vision

New initiatives fail when people feel disempowered by huge obstacles. These roadblocks need to be confronted else the change will be undermined.

- **Psychological**: Do they stem from fear, lack of knowledge or something else?
- **Structural**: Narrow organisational structures that force people to choose between the Vision and their own self-interest must be addressed
- **Active resistance**: Roadblocks that exist because people refuse to change

4.6 Error #6: Failing to Create Short-Term Wins

Complex change efforts require short-term goals in support of long-term objectives. Meeting these goals allows the team to celebrate and maintain momentum.

In the Information Management context, a short-term win is the resolution of an identified problem, for example resolving a problem of inconsistent understanding of data.

4.7 Error #7: Declaring Victory Too Soon

Kotter – changing a company can take three to ten years. The changes must be embedded in the culture.

Example in Information Technology – the implementation of a tool is the route to resolving an issue.

Example Scenario	How it might manifest
Addressing Data Quality	"We've bought a Data Quality tool. That's fixed that now." • No one in the organization is reviewing or acting on Data Quality reports
Confusing capability delivery with implementation and operation	"We've implemented the reporting stack for Regulation X. We are now compliant with the legislation." • Regulatory requirement changes • Nobody is reviewing or acting on issues identified in reporting
Migration of data	"All the data in System X is now in System Y." • Record counts match, but the data in System Y is incomplete, or truncated due to failures in the migration process. Manual interventions needed

Table 37 Declaring Victory Too Soon Scenarios

4.8 Error #8: Neglecting to Anchor Changes Firmly in the Corporate Culture

Organisations don't change, people do. Kotter is clear – ignore culture at your peril.

Keys to anchoring change into the culture of the organisation:

- Consciously show people how specific behaviours and attitudes have influenced performance
- Take time to embed the change in the next generation of management

5 Kotter's Eight Stage Process for Major Change

Kotter's obstacles to change

- Inward focussed cultures
- Paralysing bureaucracy
- Parochial politics
- Low levels of trust
- Lack of teamwork
- Arrogance
- Lack of or failure of leadership
- Fear of the unknown

Kotter's Eight step model for major change: No short cut, follow all the steps.

- First four steps: Soften entrenched status quo positions
- 5-7: Introduce new ways of working
- Last step locks the change

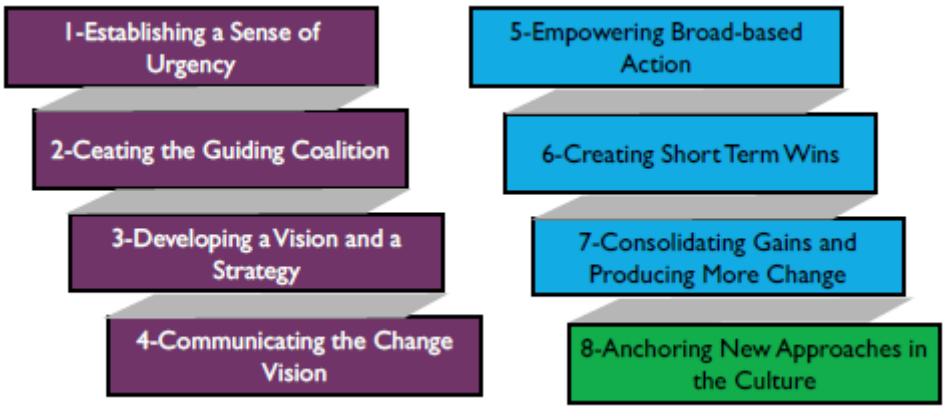

1-Establishing a Sense of Urgency
2-Ceating the Guiding Coalition
3-Developing a Vision and a Strategy
4-Communicating the Change Vision
5-Empowering Broad-based Action
6-Creating Short Term Wins
7-Consolidating Gains and Producing More Change
8-Anchoring New Approaches in the Culture

Figure 114 Kotter's Eight Stage Process for Major Change

5.1 Establishing a Sense of Urgency

People will withhold cooperation if they think something is unnecessary. Eliminate complacency and create a sense of urgency:

- Regulatory changes
- Threats to security of information
- Risks to business continuity
- Changes to business strategy
- Mergers or acquisitions
- Regulatory audit or litigation threats
- Changes to technology
- Changes of capability of competitors in the market
- Media commentary on organisation's information management issues

5.1.1 Kotter Identifies 9 Sources of Complacency:

- No visible crisis
- Trappings of success drown out urgency
- Measuring staff against low performance standards
- Narrow functional goals – no one is responsible if performance is poor
- Internal planning and control systems can be rigged to meet goals
- Performance feedback is internal and faulty – no sanity check
- Identified problems or external feedback is attacked - culture of "kill the messenger"
- People don't accept things they don't want to hear
- Happy talk from management can lead to a false sense of security

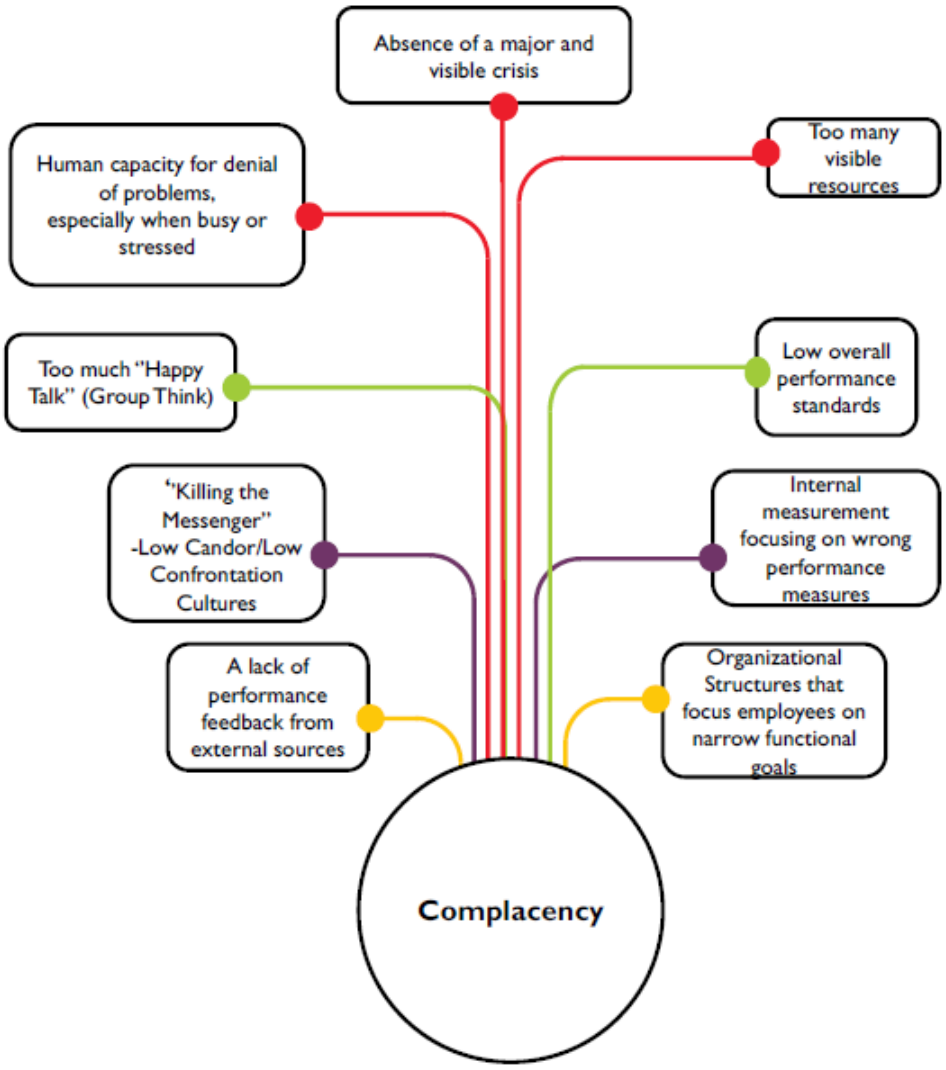

Absence of a major and visible crisis
Too many visible resources
Human capacity for denial of problems, especially when busy or stressed
Too much "Happy Talk" (Group Think)
Low overall performance standards
"Killing the Messenger" -Low Candor/Low Confrontation Cultures
Internal measurement focusing on wrong performance measures
A lack of performance feedback from external sources
Organizational Structures that focus employees on narrow functional goals
Complacency

Figure 115 Sources of Complacency

Kotter suggests the support 75% of managers is the critical mass to sustain transformation in the long term.

5.2 The Guiding Coalition

Create a Guiding Coalition to avoid the Lone CEO/ Lone Champion or Low Credibility Committee scenarios.

Four key characteristics of the Guiding Committee:

- **Position Power:** Enough key players on board
- **Expertise:** Relevant points of view represented to make informed, intelligent decisions
- **Credibility:** People with good reputations in the organisation
- **Leadership:** Enough proven leaders on board

Avoid:

- **Naysaying:** Hamper positive and open dialogue
- **Distraction:** Members must stay focusses on the change activity. Unfocussed individuals can take the team off course

- **Selfishness:** No hidden agendas

Leaders with the following personality traits, Kotter says, will derail the team:

- **Egos**: Do not let others contribute
- **Snakes:** Create and spread mistrust
- **Reluctant players:** Senior people who don't grasp the urgency for change

An effective team is built on trust and a common goal.

Combat "Group Think". Everyone goes along with a proposal even if they have reservations. Symptoms of Group Think:

- No one raises objections
- No alternatives are offered
- Different perspectives are quickly dismissed and die
- Information that might challenge thinking is not sought

Prevent Group Think by:

- Follow scientific method of gathering data
- Develop criteria for evaluating decisions
- Learn to work together
- Brainstorm
- Leaders should speak last
- Actively search for outside knowledge to bring to meetings
- Once a solution has been identified, have the team develop a Plan A and a Plan B

5.3 Developing a Vision and Strategy

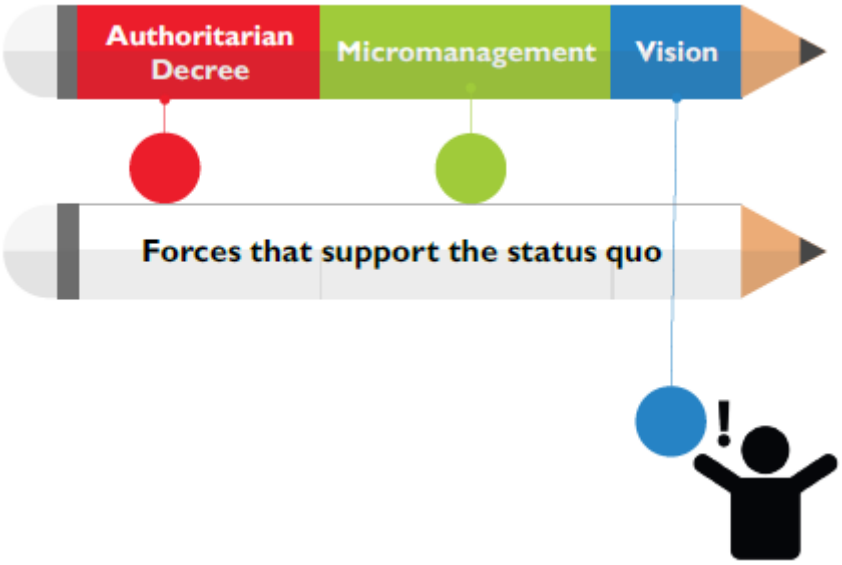


Figure 116 Vision Breaks Through Status Quo

5.3.1 Why Vision is Essential

Vision is a picture of the future with some commentary about why people should strive to create that future. A good vision shares three important purposes:

- **Clarification:** Clarifies the direction of change, sets key parameters. Clears clutter to enable the team to focus
- **Motivation:** Can give people a cause to fight for
- **Alignment:** Shared sense of direction, avoids conflict and non-stop meetings

5.3.2 The Nature of an Effective Vision

Characteristics of an effective vision:

- **Imaginable:** Conveys a picture of the future
- **Desirable:** Appeals to long-term interests of employees and other stakeholders
- **Feasible:** Realistic and attainable goals
- **Focussed:** Clear enough to provide guidance in decision making
- **Flexible:** General enough to allow for alternative plans if conditions change
- **Communicable:** Easy to share in five minutes or less

5.3.3 Creating an Effective Vision (Kotter):

- **First draft:** Single individual makes initial statement reflecting the needs and dreams
- **Role of the Guiding Coalition:** Guiding Coalition reworks the first draft to fit wider strategic perspective
- **Importance of Teamwork:** Encourage people to contribute
- **Role of the head and heart:** Both analytical thinking and dreaming
- **Messiness of the process:** Debate, rework and change is good
- **Time Frame:** Constantly evolving
- **End product:** A direction for the future that is desirable, feasible, focussed, flexible and can be conveyed in five minutes or less

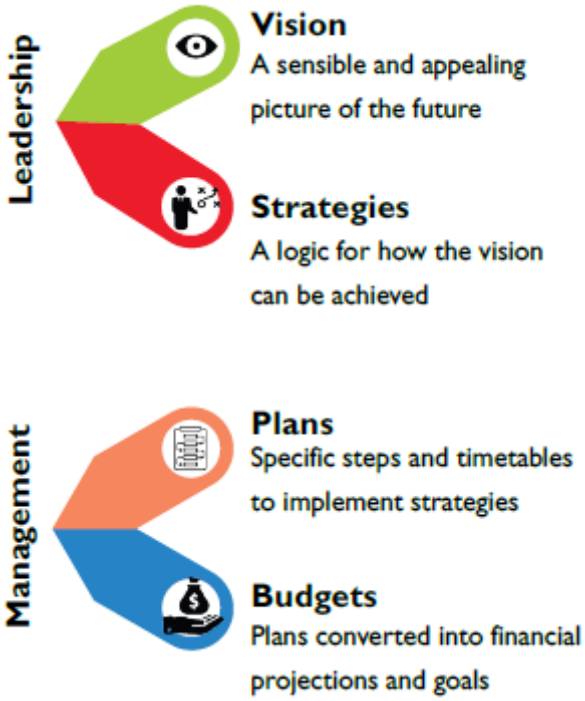


Figure 117 Management / Leadership Contrast

5.4 Communicating the Change Vision

A vision only has power when those involved have a common understanding of the goals and direction. Problems in communication:

- **Failure to communicate,** or not enough
- **Poor communication:** Cumbersome wording, people don't listen carefully
- **Not communicating far enough:** Leaders need to communicate out into broader circumstances.

Questions about the vision need to be dealt with carefully, not a FAQ dump. Information overload clouds the vision or causes panic.

Kotter identifies seven key elements in effective communication of vision:

- **Keep it simple:** Strip out jargon, internal vocabulary and complex sentences. It is hard to connect with unnatural language
- **Use metaphor, analogy and example:** A verbal or graphical picture
- **Use multiple forums:** Message needs to be communicable from elevator pitch to broadcast memo, large and small meetings. People process information differently so a multi-channel / multi-format approach to disseminate the same message will reach most
- **Repeat, repeat, repeat:** Ideas need to be heard many times to be internalised and understood
- **Lead by example:** Behaviour of important people must be consistent with vision
- **Explain seeming inconsistencies:** Loose ends undermine credibility. Inconsistencies are sometimes unavoidable. Explain why the inconsistency is valid, and how long it will exist for if it is transient. The explanation must be simple, clear and honest
- **Give and take, listen and be listened to:** Two-way communication is more powerful than one-way communication. Seek first to understand, then to be understood

6 The Formula for Change

The Gleicher Formula describes the factors that need to be in place to overcome change.

$$C = (D \times V \times F) > R$$

Change (C) occurs when the level of dissatisfaction with the status quo (D) is combined with a vision for a better alternative (V) and some actionable first steps to get there (F) and the product of the three is enough to overcome resistance (R) in the organisation.

Influencing the variables to influence the success of the change effort must be done with care.

7 Diffusion of Innovations and Sustaining Change

Understanding how new ideas spread throughout the organisation.

Diffusion of Innovations: Theory to explain how, why and at what rate new ideas and technology spreads through cultures (Everett Rogers, 1962)

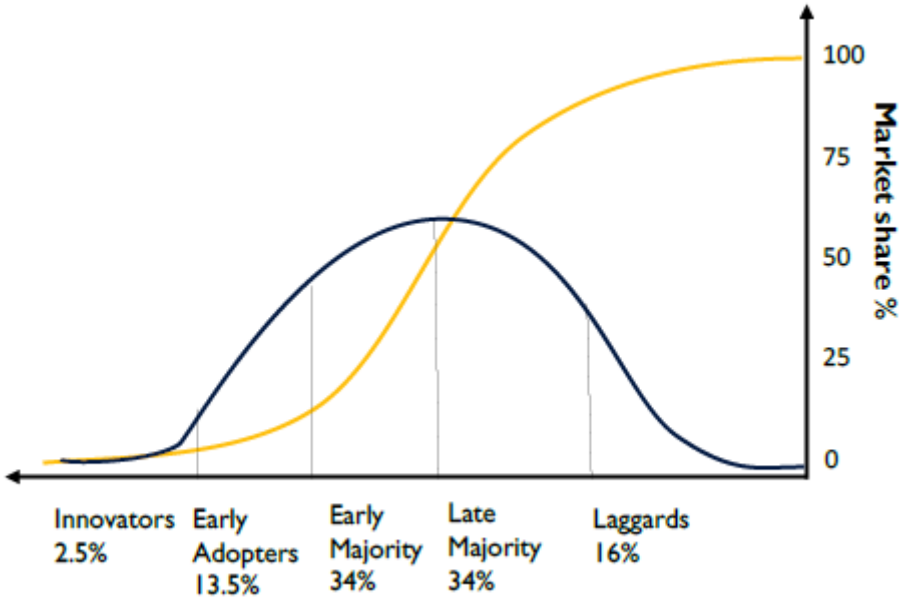


Figure 118 Everett Rogers Diffusion of Innovations

Adopter Category	Definition (Information Management Perspective)
Innovators	Innovators are the first individuals to spot a better way to tackle problems with the quality of information. They take risks trying to develop profiling of data, build tentative scorecards, and begin to put the symptoms experienced by the business into the language of Information Management. Often these innovators will use their own resources to get information and develop skills about best practices.
Early Adopters	Early Adopters are the second fastest category of individuals to adopt an innovation. These individuals have the highest degree of opinion leadership among the other adopter categories. They are perceived as 'visionary' managers (or experienced managers, or managers responsible for emergent business strategy areas) who have realized information quality issues are a barrier to their success. Often they piggy back on the initial work of the Innovators to develop their business case and begin to formalize information practices.
Early Majority	It takes the Early Majority significantly longer than the Early Adopters to adopt an innovation. Early Majority tend to be slower in the adoption process, have above average social status, contact with early adopters, and seldom hold positions of opinion leadership in a system. They could be in the 'traditional core' areas of the organization where the impact of poor quality data is masked as the 'cost of business'.
Late Majority	Individuals in the Late Majority approach an innovation with a high degree of skepticism and after most society has adopted the innovation. Late Majority typically have below average social status, very little financial lucidity, in contact with others in late majority and early majority, very little opinion leadership. In Information Management terms, these can be areas of the organization where tight budgets might combine with skepticism about the proposed changes to generate resistance.
Laggards	Laggards are the last to adopt an innovation. Individuals in this category show little to no opinion leadership. They are typically averse to change-agents and tend to be advanced in age. Laggards tend to focus on 'traditions'. In Information Management, terms these are often the people or areas of the business who resist because the 'new thing' means having to do the 'old thing' differently or not at all.

Table 38 Diffusion of Innovations Categories Adapted to Information Management

7.1 Challenges to Overcome as Innovations Spread

Two challenge areas:

- **Breaking past Early Adopter stage**: Early Adopters identify sufficient dissatisfaction with the status quo that they make and persist with the change to the 'Tipping Point', where the innovation is accepted by enough people to become mainstream
- **As the innovation moves out of Late Majority stage into the Laggards:** Accept that there will always be some resisters, and decide what to do about them

7.2 Key Elements in the Diffusion of Innovation

Four key elements:

- **Innovation:** An idea, practice or object perceived as new
- **Communication channels:** Means by which messages get from one individual to another
- **Time:** Speed at which the innovation is accepted by members of the social system
- **Social system:** The set of interrelated units engaged in joint problem solving to accomplish a common goal.

7.3 The Five Stages of Adoption

Stage	Definition
Knowledge	In the knowledge stage the individual is first exposed to an innovation but lacks information about the innovation. During this stage the individual has not yet been inspired to find more information about the innovation.
Persuasion	In the persuasion stage the individual is interested in the innovation and actively seeks information about the innovation.
Decision	In the Decision stage the individual weighs the advantages and disadvantages of using the innovation and decides whether to adopt or reject it. Rogers notes that the individualistic nature of this stage makes it the most difficult stage about which to acquire empirical evidence.
Implementation	In the Implementation stage the individual employs the innovation and determines its usefulness or searches for further information about it.
Confirmation	In the Confirmation stage, the individual finalizes his/her decision to continue using the innovation and may end up using it to its fullest potential.

Table 39 The Stages of Adoption (Adapted from Rogers, 1964)

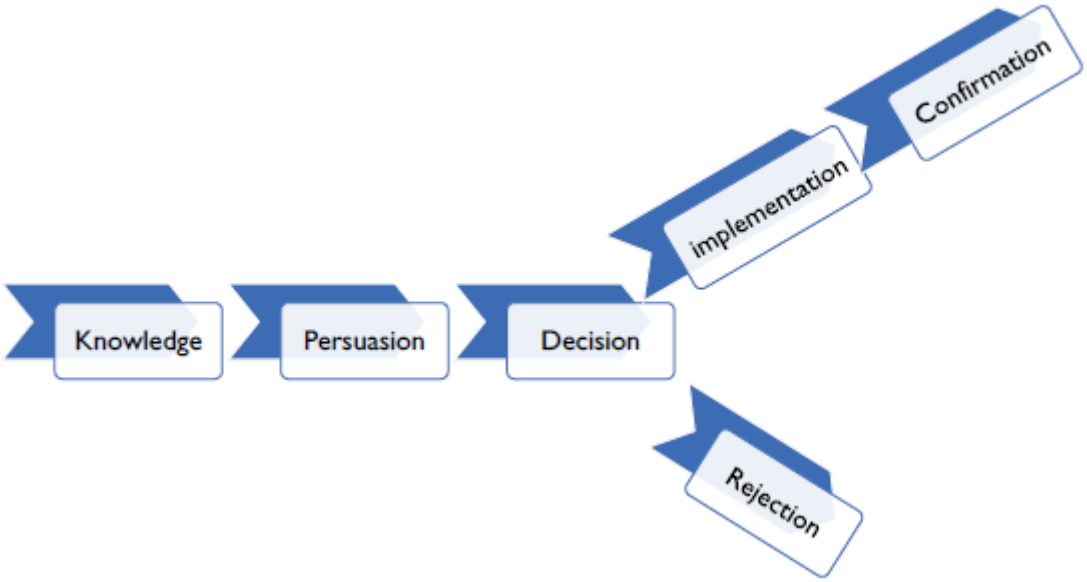


Figure 119 The Stages of Adoption

7.4 Factors Affecting Acceptance or Rejection of an Innovation or Change

Whether the innovation offers an advantage over the previous way of doing things or is compatible with their way of working. Factors considered:

- Complexity or simplicity of the change

- **Trialability**: How easy is it to experiment with the change
- **Observability:** The extent that the innovation is visible
    - Drives communication
    - Plan how to handle negative feedback

8 Sustaining Change

Starting change requires clear vision of first steps, dissatisfaction with status quo, a Guiding Coalition and a plan for Change Agents.

8.1 Sense of Urgency/Dissatisfaction

Maintain the sense of urgency.

8.2 Framing the Vision

A vision should provide direction for the organisation, allow broad based action and not create a dead end once for change leaders once the initial projects are delivered.

8.3 The Guiding Coalition

Guides and evolves change in the organisation.

8.3.1 Relative Advantage and Observability

Principles, practices and tools of a change initiative may be transferrable to other initiatives.

9 Communicating the Data Management Value

9.1 Communication Principles

Communication needs to be carefully planned to communicate the importance of data management to people who do not understand its importance. An overall communications plan, and each individual communication should:

- Have a clear objective and desired outcome
- Consist of key messages to support the desired outcome
- Be tailored to audience/stakeholders
- Be delivered via media appropriate to audience/stakeholders

General goals of communication:

- Informing
- Educating
- Setting goals or vision
- Defining a solution to a problem
- Promoting change
- Influencing or motivating action
- Gaining feedback
- Generating support

Data management communications should strive to:

- Convey the tangible and intangible value of DM initiatives
- Describe how DM capabilities contribute to business strategy and results
- Share examples of how DM reduces cost and risk, and supports revenue and decision quality

- Increase base of knowledge about DM in the organisation

9.2 Audience Evaluation and Preparation

Identify audience through stakeholder analysis, then tailor communication.

Getting people to see how their interest aligns with the program persuades them to act on communications:

- Solve problems
- Address pain points
- Present changes as improvements
- Have a vision of success
- Avoid jargon
- Share stories and examples
- Recognise fear as motivation

9.3 The Human Element

As people are influenced by their colleagues and leaders, communication should use a stakeholder analysis to identify groups with like interests and needs.

9.4 Communication plan

Element	Description
Message	The information that needs to be conveyed.
Goal / Objective	The desired outcome of conveying a message or set of messages (i.e., why the message needs to be conveyed).
Audience	Group or individual targeted by the communication. The plan will have different objectives for different audiences.
Style	Both the level of formality and the level of detail in messages should be tailored to the audience. Executives need less detail than teams responsible for implementation of projects. Style is also influenced by organizational culture.
Channel, Method, Medium	The means and format through which the message will be conveyed (e.g., web page, blog, email, one-on-one meetings, small group or large group presentations, lunch and learn sessions, workshops, etc.) Different media have different effects.
Timing	How a message is received may be influenced by when it is received. Employees are more likely to read an email that comes out first thing Monday morning than one that comes out last thing on Friday afternoon. If the purpose of a communication is to gain support in advance to a budget cycle, then it should be timed in relation to the budget cycle. Information about impending changes to processes should be shared in a timely manner and in advance of a change taking place.
Frequency	Most messages need to be repeated in order to ensure all stakeholders hear them. The communications plan should schedule the sharing of messages so that repetition is helpful in getting the message across and does not become an annoyance. In addition, ongoing communications (for example, a newsletter) should be published based on an agreed-to schedule.
Materials	The communications plan should identify any materials that need to be created to execute the plan. For example, short and long versions of presentation and other written communications, elevator speeches, executive summaries, and marketing materials like posters, mugs, and other means of visual branding.
Communicators	The communications plan should identify the person or people who will deliver communications. Often the person delivering the message has a profound influence on the target audience. If the data management sponsor or other executive delivers a message, stakeholders will have a different response than if a lower level manager delivers it. Decisions about who will communicate which messages to which stakeholders should be based on the goals of the message.
Expected Response	The communications plan should anticipate how different stakeholder groups, and sometimes how individual stakeholders, will respond to communications. This work can be accomplished by anticipating questions or objections and formulating responses. Thinking through potential responses is a good way to clarify goals and build robust messages to support them.
Metrics	The communications plan should include measures of its own effectiveness. The goal is to ensure that people have understood and are willing and able to act on the messages in the plan. This can be accomplished through surveys, interviews, focus groups, and other feedback mechanisms. Changes in behavior are the ultimate test of a communications plan's success.
Budget and Resource Plan	The communications plan must account for what resources are needed to carry out goals within a given budget.

Table 40 Communications Plan Elements

9.5 Keep Communicating

A DM program is an ongoing effort which needs to be sustained through communication. Employees and roles change, and the competition for funding remains. Successful communication provides better understanding of the way that DM can generate business value from information assets and have a long-lasting impact on the organisation.

10 Metrics (Not in Revised Version)

- Measure Stakeholder contributions
- Measure target culture progress
- Check current behaviour, process and system

- Adoption and Adoption Improvement Rate
- DM Influence
- Process Efficiency & Effectiveness
- Data Risk Management
- DM & Business Innovation
- Trusted Analytics
- Effectiveness of training and Communication interventions

11 DM Org and Change Management and Governance

NOTE: This section is not in the DMBOK. Added by Howard Diesel in his Data Governance training.

- Communication: People get Data (and data jokes)
    - Data Strategy Vision
    - Knowledge Area Vision
- Data Culture Norms: Sophistication Movement
    - Disabled, Enabled, Led, Data-Driven & Knowledge-Driven
- Structure Growth: Levels of Excitement Vs Complacency
    - Committees
    - Teams
    - Individuals
- Knowledge Management
    - Sustainable Change & Growth

Appendix 1
DMBOK Version 1 Notes

This document is your reference in the event of a question based on the DMBOK Version 1. Sample questions are listed with their relevant DMBOK V1 section.

In most cases, DMBOK V1 questions refer to "The DAMA-DMBOK". DMBOK Version 2 questions usually refer to "The DAMA-DMBOK Version 2 or V2".

DMBOK 1 sample questions and their relevant sections are listed here.

1 Data Management

1.1 According to the DAMA DMBoK, what parts of the Data Lifecycle are integral parts of the SDLC?

Correct: **Plan, Specify, Enable**

1.2 The DMBoK identifies which of the following as common stages in the lifecycle of the information asset?

Correct: **Plan, Specify, Enable, Create and Acquire, Maintain & Use, Archive & Retrieve, Purge**

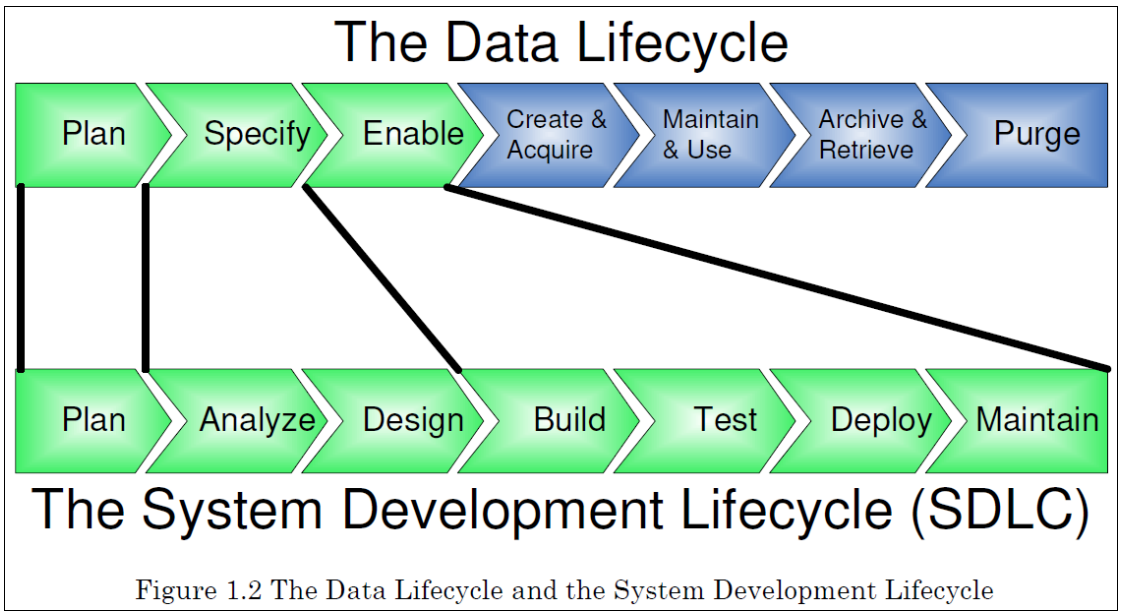


Figure 1.2 The Data Lifecycle and the System Development Lifecycle

According to the DMBoK, which is not a component of a Data Management strategy?

Correct: **Identifying individuals for Data Management roles**

This is a Version 2 question: Section 2.6 page 31

1.3 According to the DMBOK, which of the following are included as Data Management goals?

Correct: **We should ensure that data can be used effectively to add value to the enterprise.**

Version 2 Question: Section 1.2 Goals page 18

1.4 In the data lifecycle, the DMBoK recommends minimizing ROT data. ROT stands for:

Correct: **Data that is Redundant, Obsolete, Trivial**

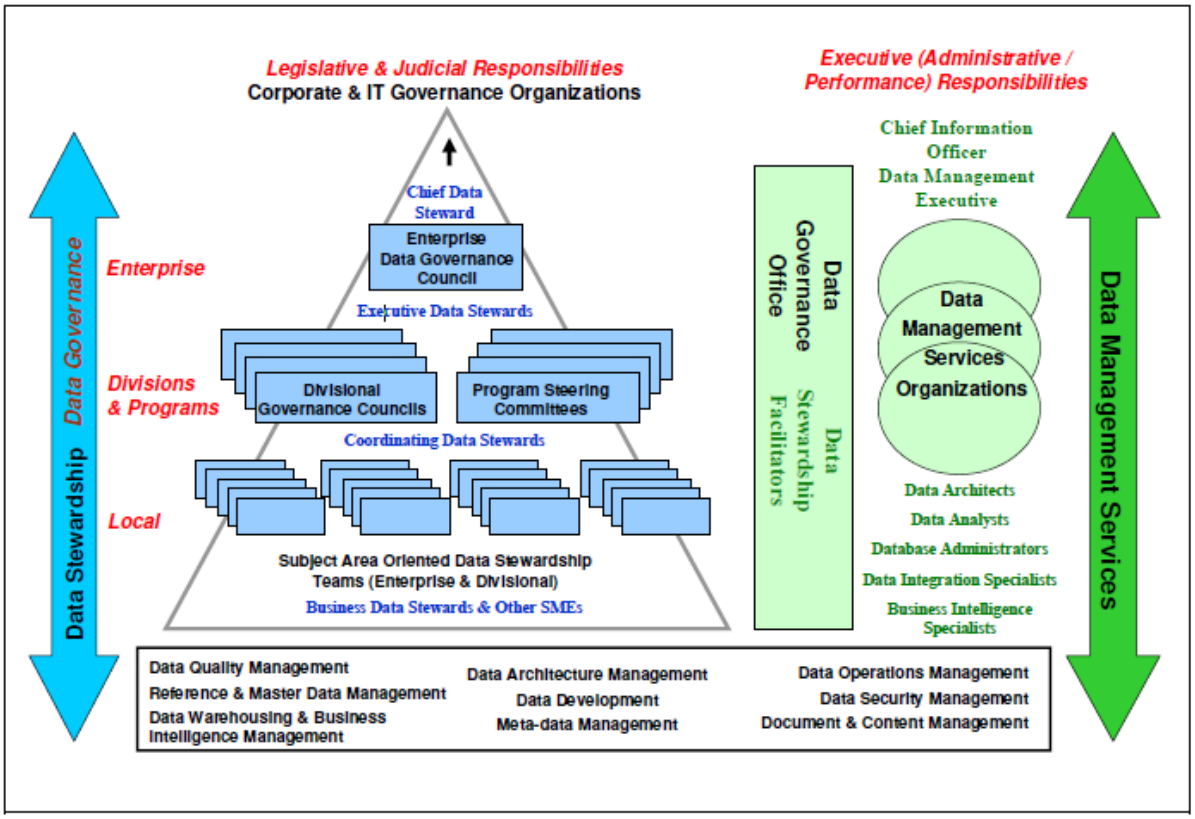
Version 2 question: Section 2.5.9 page 29

- **Data Management efforts should focus on the most critical data**: Organizations produce a lot of data, a large portion of which is never actually used. Trying to manage every piece of data is not possible. Lifecycle management requires focusing on an organization's most critical data and minimizing data ROT (Data that is Redundant, Obsolete, Trivial) (Aiken, 2014).

2 Data Governance

2.1 According to the DAMA DMBoK, the Data Governance Council (DGC) is the highest-authority organization for Data Governance in an organization. Who should typically chair this Council?

Correct Answer: **Chief Data Steward (Business)/Chief Data Officer**



3 Data Architecture

3.1 DMBOK Version 1 identifies the SAM and the highest level model. The second layers are conceptual views of each subject area.

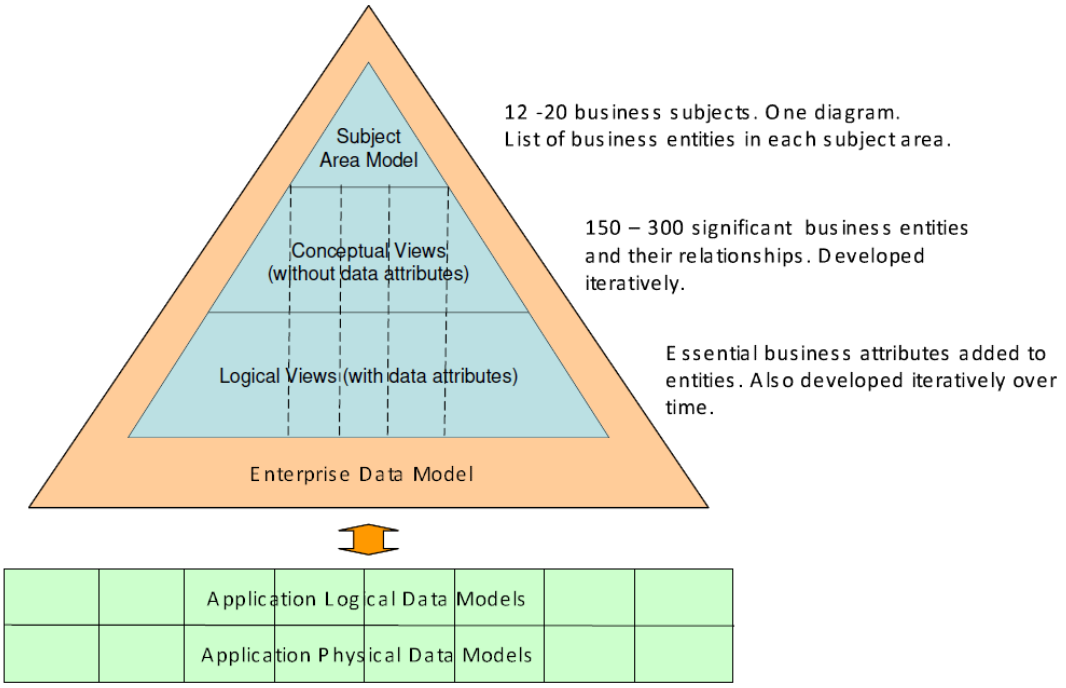

Subject Area Model
Conceptual Views (without data attributes)
Logical Views (with data attributes)
Enterprise Data Model
12 -20 business subjects. One diagram. List of business entities in each subject area.
150 – 300 significant business entities and their relationships. Developed iteratively.
Essential business attributes added to entities. Also developed iteratively over time.
Application Logical Data Models
Application Physical Data Models

Figure 4.3 Enterprise Data Model Layers

4 Data Modelling and Design

4.1 What is the difference between cardinality rules and data integrity rules?

Correct: **Cardinality rules define the quantity of each entity instance that can participate in a relationship between two entities, and referential integrity rules ensure valid values**

4.2 What is the definition of cardinality?

Correct: **Defines how many instances of one entity are related to instances of another entity**

Data models express two primary types of data rules:

Cardinality rules define the quantity of each entity instance that can participate in a relationship between entities. For example, “Each company can employ many persons.”

Referential Integrity rules ensure valid values. For example, “A person can exist without working for a company, but a company cannot exist unless at least one person is employed by that company.”

Levels of Data Models from the DMBOK 1:

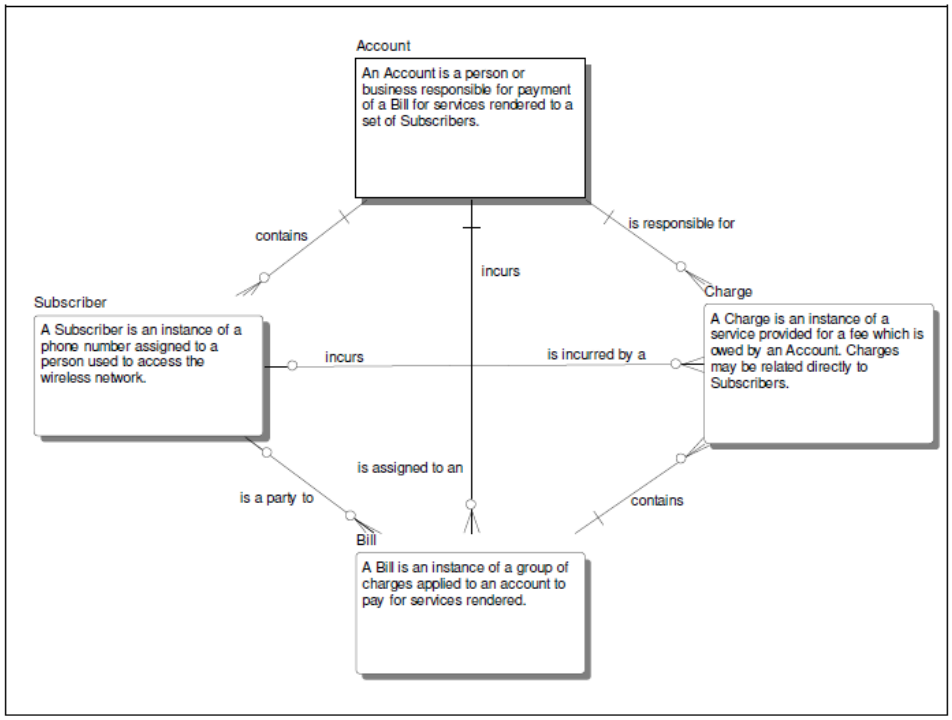

Account
An Account is a person or business responsible for payment of a Bill for services rendered to a set of Subscribers.
contains
incurs
is responsible for
Subscriber
A Subscriber is an instance of a phone number assigned to a person used to access the wireless network.
incurs
is incurred by a
Charge
A Charge is an instance of a service provided for a fee which is owed by an Account. Charges may be related directly to Subscribers.
is assigned to an
is a party to
contains
Bill
A Bill is an instance of a group of charges applied to an account to pay for services rendered.

Figure 5.3 Conceptual Data Model Example

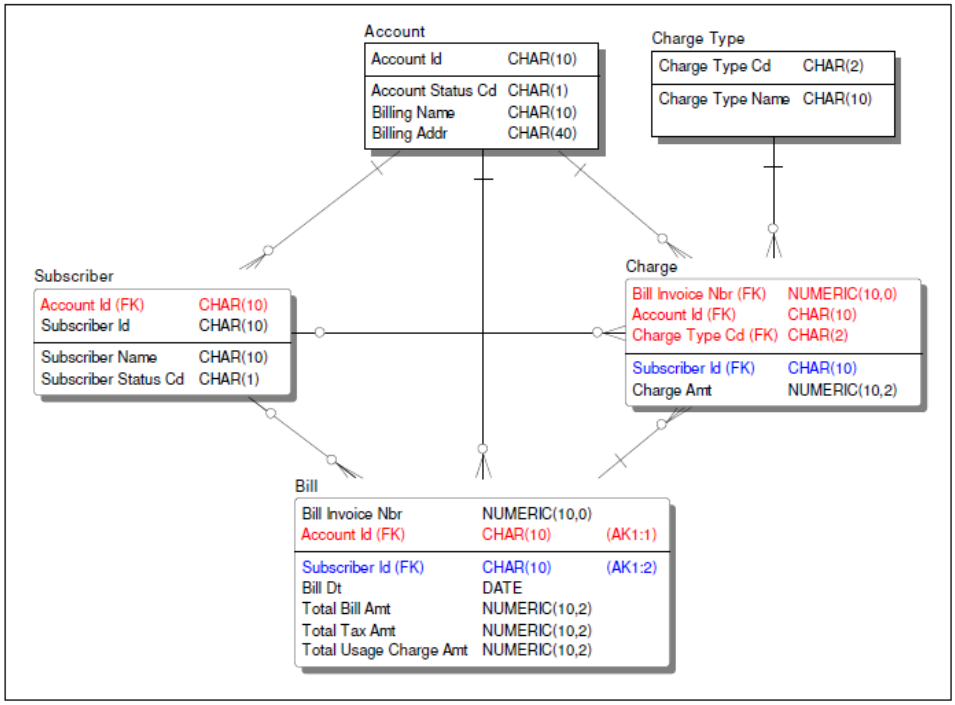


Figure 5.4 Logical Data Model Example

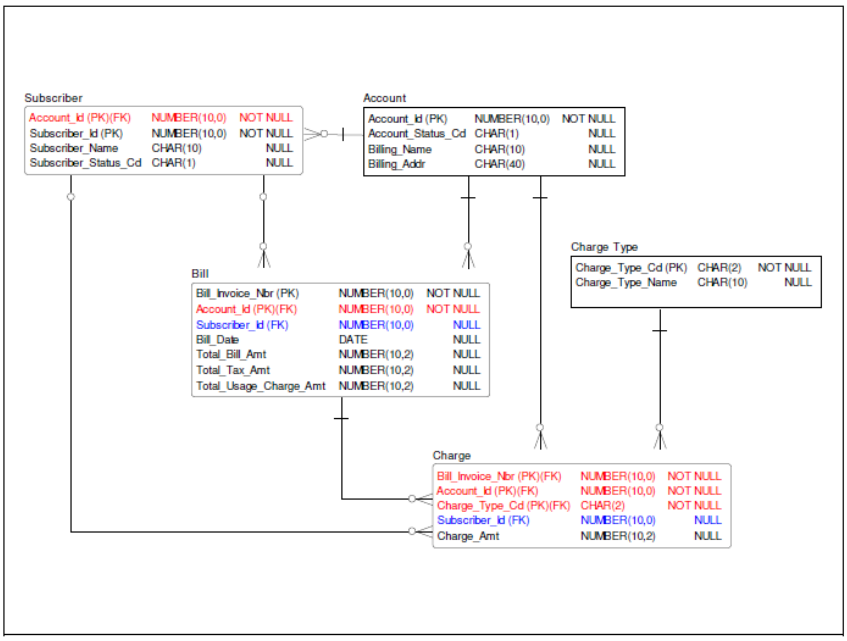


Figure 5.5 Physical Data Model Example

Denormalization is the deliberate transformation of a normalized logical data model into tables with redundant data. In other words, it intentionally puts one data element in multiple places. This process does introduce risk of data errors due to duplication. Implement data quality checks to ensure that the copies of the data elements stay correctly stored. Only denormalized specifically to **improve database query performance**, by either segregating or combining data to reduce query set sizes, combining data to reduce joins, or performing and storing costly data calculations.

5 Document and Content Management

5.1 Non value-added information is often not removed because:

Correct: **The policies are unclear of what is defined as non-value-added, storage is cheap so there is no cost driver, and it takes more effort to dispose than to keep.**

5.2 Which of the following is a reason why organizations do not dispose of non-value-adding information?

Correct: **Storage is cheap and easily expanded**

Non-value added information should be removed to avoid wasting space and the costs of maintenance. Develop policies and procedures.

Many organisations do not prioritise removing non-value added information because:

- Policies are not adequate:
  - One person's non-value added information may be valuable to another person
  - There may be possible future needs for the information
- There is no buy-in for records management
  - Inability to decide what to keep
  - Perceived cost of making a decision to remove the information
  - Electronic space is cheap. it is easier to buy more than to put archive/disposal processes in place

6 Data Warehousing and BI

6.1 A comparatively new architectural approach is where volatile data is provisioned in a data warehouse structure to provide transactional systems with a combination of historical and near real time data to meet customer needs. This is a definition of:

Correct: **Active Data Warehousing**

Active Data Warehousing

With the onset of Operational BI (and other general requirements from the business) pushing for lower latency and more integration of real time or near real time data into the data warehouse, new architectural approaches are emerging to deal with the inclusion of volatile data. A common application of operational BI is the automated banking machine (ABM) data provisioning. When making a banking transaction, historical balances and new balances resulting from immediate banking actions, need to be presented to the banking customer real-time. The impact of the changes from new volatile data must be isolated from the bulk of the historical, non-volatile DW data. Typical architectural approaches for isolation include a combination of building partitions and using union queries for the different partitions, when necessary.

Fact Tables
Fact tables contain important business measures. They are entities which contain attributes representing measures (facts). Rows correspond to a particular measurement and are numeric. Fact tables have many rows and take up the most space in the database (around 90%).

Dimension Tables
Represent the important objects of the business and contain textual descriptions of the business. They are the source of the "query by" or "report by" constraints. They are denormalised and account for 10% of the total data. All designs will have a Date dimension and Organisation or Party Dimension as a minimum. Dimensions have surrogate or natural key.

Figure 9.5 relates the classes of BI tools to the respective classes of BI users for those tools.

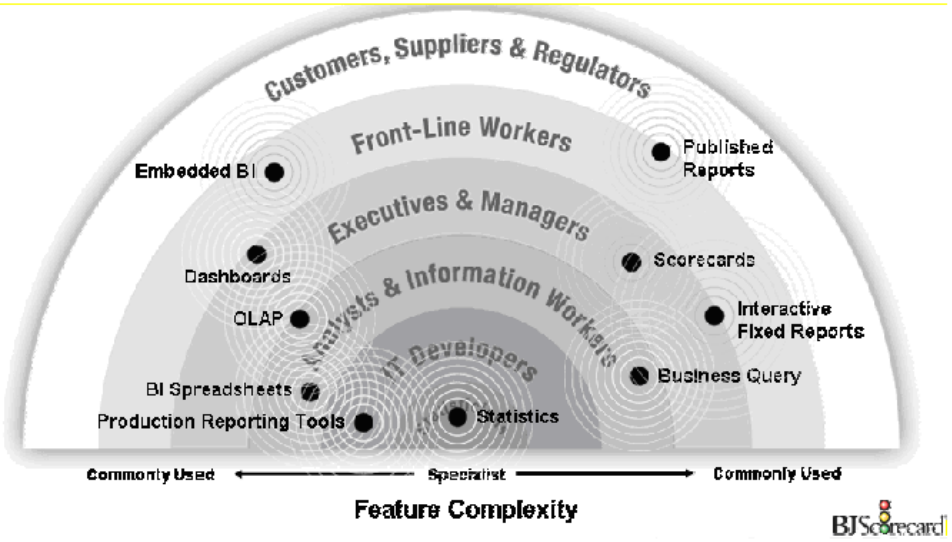


Figure 9.5 What BI Tools for What Users?

7 Metadata Management

Industry / Consensus and International Meta-data Standards:

NOTE: This question is incorrect, they are not the same, but this is the answer in the exam.

7.1 What is the difference between an industry and a consensus meta-data standard?

Correct: **The terms are used interchangeably to describe the same concept.**

Two major types of meta-data standards exist:
- industry or consensus standards (terms used interchangeably)
- international standards.

Metadata Type Questions:

7.2 What type of Meta-Data provides developers and administrators with knowledge and information about systems?

Correct: **Technical Operational Meta-Data**.

7.3 These are examples of which type of Meta-Data: Data Stores & Data Involved, Government/ Regulatory Bodies; Roles & Responsibilities; Process Dependencies and Decomposition:

Correct: **Process Meta-Data**

NOTE: DMBOK 1 and DMBOK 2 have different lists of metadata types:

DMBOK V2 valid types are: **Business**, **Technical** and **Operational**.

Look up DMBOK V 2 Section 1.3.2 page 422 to 424.

Metadata types according to DMBOK 1

Meta-data is classified into **four** major types: **business, technical and operational, process,** and **data stewardship**.

Business Metadata:

Business meta-data includes the business names and definitions of subject and concept areas, entities, and attributes; attribute data types and other attribute properties; range descriptions; calculations; algorithms and business rules; and valid domain values and their definitions. **Business meta-data relates the business perspective to the meta-data user**.

- Business data definitions, including calculations.
- Business rules and algorithms, including hierarchies.
- Data lineage and impact analysis.
- Data model: enterprise level conceptual and logical.

- Data quality statements, such as confidence and completeness indicators.
- Data stewardship information and owning organization(s).
- Data update cycle.
- Historical data availability.
- Historical or alternate business definitions.
- Regulatory or contractual constraints.
- Reports lists and data contents.
- System of record for data elements.
- Valid value constraints (sample or list).

Technical and operational meta-data:

Technical and operational meta-data **provides developers and technical users with information about their systems**.

Technical meta-data includes physical database table and column names, column properties, other database object properties, and data storage. The database administrator needs to know users patterns of access, frequency, and report / query execution time. Capture this meta-data using routines within a DBMS or other software.

Operational meta-data is targeted to the IT operations users' needs, including information about data movement, source and target systems, batch programs, job frequency, schedule anomalies, recovery and backup information, archive rules, and usage.

Examples of technical and operational meta-data include:

- Audit controls and balancing information.
- Data archiving and retention rules.
- Encoding / reference table conversions.
- History of extracts and results.
- Identification of source system fields.
- Mappings, transformations, and statistics from the system of record to target
- data stores (OLTP, OLAP).
- Physical data model, including data table names, keys, and indexes.
- Program job dependencies and schedule.
- Program names and descriptions.
- Purge criteria.
- Recovery and backup rules.
- Relationships between the data models and the data warehouse / marts.
- Systems of record feeding target data stores (OLTP, OLAP, SOA).
- User report and query access patterns, frequency, and execution time.
- Version maintenance.

Process Metadata

Process meta-data is data that defines and describes the characteristics of other system elements (processes, business rules, programs, jobs, tools, etc.).

Examples of process meta-data include:

- Data stores and data involved.
- Government / regulatory bodies.
- Organization owners and stakeholders.
- Process dependencies and decomposition.

- Process feedback loop documentation.
- Process name.
- Process order and timing.
- Process variations due to input or timing.
- Roles and responsibilities.
- Value chain activities.

Data Stewardship Metadata

Data stewardship meta-data is data about data stewards, stewardship processes, and responsibility assignments. Data stewards assure that data and meta-data are accurate, with high quality across the enterprise. They establish and monitor sharing of data.
Examples of data stewardship meta-data include:

- Business drivers / goals.
- Data CRUD rules.
- Data definitions - business and technical
- Data owners.
- Data sharing rules and agreements / contracts.
- Data stewards, roles and responsibilities.
- Data stores and systems involved.
- Data subject areas.
- Data users.
- Government / regulatory bodies.
- Governance organization structure and responsibilities

Unstructured Metadata:

Examples of descriptive meta-data include:

- Catalog information.
- Thesauri keyword terms.

Examples of structural meta-data include:

- Dublin Core.
- Field structures.
- Format (Audio / visual, booklet).
- Thesauri keyword labels.
- XML schemas.

Examples of administrative meta-data include:

- Source(s).
- Integration / update schedule.
- Access rights.
- Page relationships (e.g. site navigational design).

Bibliographic meta-data, record-keeping meta-data, and preservation meta-data are all meta-data schemes applied to documents, but from different focuses:

- *Bibliographic meta-data* is the library card of the document.
- *Record-keeping meta-data* is concerned with validity and retention.
- *Preservation meta-data* is concerned with storage, archival, condition, and conservation of material

Meta-data is a broad term that includes many potential subject areas. These subject areas include:

- **Business analytics:** Data definitions, reports, users, usage, performance.
- **Business architecture:** Roles and organizations, goals and objectives.
- **Business definitions:** The business terms and explanations for a particular concept, fact, or other item found in an organization.
- **Business rules:** Standard calculations and derivation methods.
- **Data governance**: Policies, standards, procedures, programs, roles, organizations, stewardship assignments.
- **Data integration:** Sources, targets, transformations, lineage, ETL workflows, EAI, EII, migration / conversion.
- **Data quality**: Defects, metrics, ratings.
- **Document content management:** Unstructured data, documents, taxonomies, ontologies, name sets, legal discovery, search engine indexes.
- **Information technology infrastructure:** Platforms, networks, configurations, licenses.
- **Logical data models**: Entities, attributes, relationships and rules, business names and definitions.
- **Physical data models:** Files, tables, columns, views, business definitions, indexes, usage, performance, change management.
- **Process models**: Functions, activities, roles, inputs / outputs, workflow, business rules, timing, stores.
- **Systems portfolio and IT governance:** Databases, applications, projects and programs, integration roadmap, change management.
- **Service-oriented architecture (SOA) information:** Components, services, messages, master data. 15. System design and development: Requirements, designs and test plans, impact.
- **Systems management:** Data security, licenses, configuration, reliability, service levels.

8 Data Quality Management

8.1 Top-down and Bottom-up data analysis and profiling are best done in concert because:

Correct: **It balances business relevance and the actual state of the data**

Prior to defining data quality metrics, it is crucial to perform an assessment of the data using two different approaches, **bottom-up** and **top-down**.

The **bottom-up** assessment of existing data quality issues involves inspection and evaluation of the data sets themselves. Direct data analysis will reveal potential data anomalies that should be brought to the attention of subject matter experts for validation and analysis. **Bottom-up** approaches highlight potential issues based on the results of automated processes, such as frequency analysis, duplicate analysis, cross-data set dependency, orphan child data rows, and redundancy analysis.

However, potential anomalies, and even true data flaws may not be relevant within the business context unless vetted with the constituency of data consumers. The **top-down** approach to data quality assessment involves engaging business users to document their business processes and the corresponding critical data dependencies. The **top-down** approach involves understanding how their processes consume data, and which data elements are critical to the success of the business application. By reviewing the types of reported, documented, and diagnosed data flaws, the data quality analyst can assess the kinds of business impacts that are associated with data issues.

8.2 The Data Quality Management cycle has four stages. Three are Plan, Monitor, and Act. What is the fourth stage?

Correct: **Deploy**

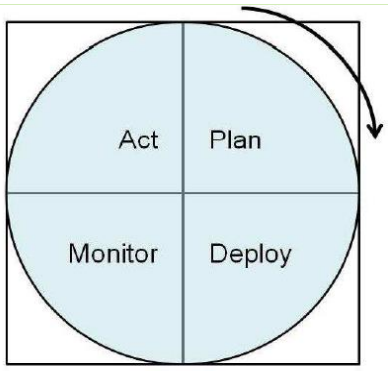


Figure 12.2 The Data Quality Management Cycle.

Appendix 2
Links to Resources

Chapter	Title	Link
	Dama South Africa Meetup group (Webinars)	https://www.meetup.com/DAMA-SA-Data-Management-Meetup/events/
	DAMA International CDMP website	https://cdmp.info
	Google drive of resources to implement NDMO Compliance	https://drive.google.com/drive/folders/1icBUW-4RPc9G9Q6frdTTuVY_PpG29zTg?usp=share_link
3	What is Data Governance?	https://www.intricity.com/learningcenter/data-governance/what-is-data-governance
4	Data Rchitecture as a water bottling plant	What is a Data Architecture? Modern Data Architectures Explained - YouTube
5	Dimensions and Measures	https://www.youtube.com/watch?v=qkJOace9FZg
5	How to read a data model	https://www.youtube.com/watch?v=adYohKb47f8
6	Database Objects	https://youtu.be/fXpjwSFgoV0
6	What is NoSQL?	https://youtu.be/BgQFJ_UNIgw
6	What is a columnar database?	https://www.youtube.com/watch?v=8KGVFB3kVHQ
6	What is a graph database?	https://www.youtube.com/watch?v=dGHSKpx4Xjs
12	What is Metadata?	https://www.youtube.com/watch?v=HXAstVP3-y0
12	Metadata Management	https://www.youtube.com/watch?v=Ioxr2mcOLjE
11	What is BI	https://www.intricity.com/learningcenter/data-warehousing/what-is-business-intelligence
11	What is a DW?	https://www.youtube.com/watch?v=KGHbY_Sales
11	Difference between Database and DW	https://www.youtube.com/watch?v=FxQG65OhXAQ
11	OLAP	https://www.youtube.com/watch?v=2ryG3Jy6eIY
13	The impact of one bad value	https://www.youtube.com/watch?v=k4gj_RdtKCw
14	Map Reduce, Hadoop and Big Data	https://www.youtube.com/watch?v=9s-vSeWej1U&list=PLIBwmCV8YuLNyzmqlDqz35BYNaIEvPNPE&index=1
14	Scaling Hadoop	https://www.youtube.com/watch?v=h5vAj9FPl0I&list=PLIBwmCV8YuLNyzmqlDqz35BYNaIEvPNPE&index=2
14	What is a Data Lake?	https://www.youtube.com/watch?v=v3yv88h68GY&list=PLIBwmCV8YuLNyzmqlDqz35BYNaIEvPNPE&index=3
14	Data Lake vs Data Warehouse	https://www.youtube.com/watch?v=AwbKwcw7bgg&list=PLIBwmCV8YuLNyzmqlDqz35BYNaIEvPNPE&index=10
14	A Periodic Table of Visualisation Methods	http://bit.ly/IX1bvI
14	Weapons of Math Destruction	https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815